Филология: научные исследования
Правильная ссылка на статью:
Сафина З.М., Лукахина Д.Н. Лексико-семантический анализ оригинала и перевода художественного текста с использованием инструментов Python // Филология: научные исследования. 2025. № 12. DOI: 10.7256/2454-0749.2025.12.76918 EDN: VBLFUI URL: https://nbpublish.com/library_read_article.php?id=76918

Лексико-семантический анализ оригинала и перевода художественного текста с использованием инструментов Python

Сафина Зарема Миниаминовна

ORCID: 0009-0009-3486-7757

кандидат филологических наук

доцент; кафедра лингводидактики и переводоведения; ФГБОУ ВО Уфимский университет науки и технологий

450076, Россия, респ. Башкортостан, г. Уфа, ул. Коммунистическая, д. 22

safinazarem@yandex.ru

Лукахина Дарья Николаевна

магистр; кафедра лингводидактики и переводоведения; ФГБОУ ВО Уфимский университет науки и технологий

450019, Россия, респ. Башкортостан, г. Уфа, Ленинский р-н, ул. Уршакская, д. 2а

dlukaxina@mail.ru

Статья из рубрики "Автоматическая обработка языка"

DOI:

10.7256/2454-0749.2025.12.76918

EDN:

VBLFUI

Дата направления статьи в редакцию:

24-11-2025

Аннотация: В статье рассматриваются возможности автоматизированного лексико-семантического анализа художественного текста и его перевода с применением библиотеки NLTK на языке программирования Python. Языки программирования позволяют значительно ускорить работу лингвистов, систематизировать результаты сбора данных в упорядоченном виде. Объектом анализа выступают оригинал повести

американского писателя XIX века Уильяма Гилмора Симмса «Grayling: or, "Murder Will Out"» и перевод на русский язык, выполненный М.Л. Павлычевой. Особое внимание уделяется выявлению различий в структуре лексики, частотности употребления слов и словосочетаний, в распределении частей речи, а также степени лексического разнообразия. Исследование направлено на выявление переводческих трансформаций, влияющих на семантическую и стилистическую организацию текста, и на оценку возможностей автоматизированного анализа при сравнении оригинала и перевода В работе применяются методы обработки естественного языка (NLP) с использованием библиотеки NLTK в среде Python, включая нормализацию текста, частеречное тегирование, частотный анализ, биграммное моделирование и расчет коэффициента лексического разнообразия. Исследование демонстрирует, что автоматизированный лексико-семантический анализ позволяет объективно выявить ключевые различия между оригиналом и переводом: в тексте перевода наблюдается более высокое лексическое разнообразие, обусловленное флективной природой русского языка и активным использованием переводческих трансформаций; увеличивается частотность связующих элементов, а тематически маркированные биграммы заменяются более нейтральными конструкциями. Кроме того, выявлены существенные ограничения стандартных NLP-инструментов при обработке русскоязычных текстов, что подчеркивает необходимость адаптации методов к специфике русского языка. Исследование подтверждает необходимость комплексного подхода при анализе оригинала и перевода художественного текста, сочетающего вычислительные методы и лингвистическую интерпретацию. Перспективы исследования включают применение современных морфологических анализаторов, расширение корпуса текстов и интеграцию методов машинного обучения для глубокого сравнительного анализа оригиналов и переводов художественных текстов.

Ключевые слова:

обработка естественного языка, Python, NLTK, лексико-семантический анализ, частотный анализ, биграммы, лексическое разнообразие, частеречное тегирование, художественный перевод, переводческие трансформации

Введение

Современные цифровые технологии кардинально изменили подходы к лингвистическим исследованиям. Особенно актуальным стало применение методов обработки естественного языка (Natural Language Processing, NLP) для автоматизированного анализа больших текстовых корпусов. NLP объединяет достижения лингвистики, информатики и искусственного интеллекта, позволяя извлекать смысловую и структурную информацию из текстов, недоступную при традиционном ручном анализе [1]. В контексте сравнительного анализа оригинала и перевода такие методы открывают новые горизонты для изучения переводческих стратегий, лексической плотности, стилистических особенностей и семантической эквивалентности [2]. Исследование с применением методов NLP «является ценным инструментом для изучения перевода художественного текста, позволяющим получить объективные данные и выявить объективные закономерности» [3, с. 64]. Цель настоящей статьи — продемонстрировать возможности библиотеки NLTK (Natural Language Toolkit) в рамках языка программирования Python для проведения лексико-семантического анализа художественного текста и его перевода. Python — многоцелевой высокоуровневый язык программирования, в котором

используется динамическая система типов и автоматическое управление памятью, а также обширная библиотека [4, p. 1856]. Применение библиотеки NLTK дает возможность осуществить предварительную обработку данных, «позволяющую машинным алгоритмам работать с текстовыми данными и выполнять анализ текста» [5, с. 177].

Обработка естественного языка является одной из высокоприоритетных задач современной лингвистики, что во многом связано с необходимостью получения быстрых и качественных результатов анализа большого массива данных. Языки программирования позволяют значительно ускорить работу лингвистов, систематизировать результаты сбора данных в упорядоченном виде, а также, в отличие от традиционных методов, могут гарантировать точность и объективность полученных результатов. Однако как в зарубежной, так и в отечественной лингвистике исследования текстов с помощью инструментов Python начали появляться относительно недавно и не носят систематический характер. Так, работа российского исследователя М. И. Ладушиной посвящена описанию техники предварительного анализа текста в лексикографических целях [6]. С. Н. Гагарин предлагает использовать язык Python и библиотеку NLTK для изучения языковых картин мира на примере парламентского дискурса [7]. Зарубежные исследователи активно разрабатывают методики использования Python для анализа текста в различных научных областях, но непосредственный лингвистический анализ текста также не представлен должным образом. Отметим проект “Digital Dostoyevsky”, в рамках которого создан корпус из пяти романов и двух повестей Ф. М. Достоевского. Текстовые разметки выполнены с помощью инструментов Python. Таким образом, лексико-семантический анализ оригинала и перевода художественного текста, направленный на изучение лексического состава текста, семантических связей между словами, частотности употребления и функциональной роли лексических единиц в дискурсе, не получил своего должного внимания в работах исследователей, что подчеркивает новизну и актуальность данной работы.

Для проведения настоящего анализа использовалась библиотека NLTK версии 3.8 на языке Python 3.11. Материалом исследования послужило произведение американского писателя XIX века У. Г. Симмса «Grayling: or, “Murder Will Out”»[8] и его русский перевод, выполненный М. Л. Павлычевой [9].

Основная часть

Ключевым понятием в NLP является нормализация текста — процесс приведения текста к унифицированному виду, пригодному для дальнейшего анализа. Он включает удаление пунктуации, приведение слов к нижнему регистру, фильтрацию стоп-слов (служебных слов, не несущих смысловой нагрузки) и лемматизацию [10]. На первом этапе для обоих текстов выполнена следующая последовательность преобразований: произведена сегментация на токены (слова), удалены все лишние символы (знаки препинания, цифры), все слова приведены к нижнему регистру (это делается для того, «чтобы одно и тоже слово, написанное строчной и прописной буквой, не воспринималось системой как два разных токена» [5, с. 178]), а также удалены стоп-слова с использованием встроенных списков NLTK для английского и русского языков (стоп-слова – единицы, не имеющие самостоятельного значения и способные повлиять на статистику встречаемости значимых слов). В результате нормализации получены тексты оригинала и перевода, подготовленные для последующего анализа, длина которых составила 9618 и 10811 токенов, соответственно.

Далее было выполнено частеречное тегирование (Part-of-Speech tagging, POS-tagging), представляющее собой «способ автоматического морфологического анализа, когда каждая словоформа входной фразы рассматривается изолировано, вне связей с другими словами предложения. В результате такого анализа каждому слову в тексте (корпусе) приписывается метка или тег, обозначающие часть речи и грамматические характеристики данного слова» [11, с. 64]. Для английского языка эта задача решена достаточно эффективно благодаря развитым морфологическим и синтаксическим моделям. Однако для флективных языков, таких как русский, POS-теггеры часто дают ошибки без использования дополнительных морфологических анализаторов (например, pymorphy3). Для частеречного тегирования использована функция `*nltk.pos_tag()*` для обеих версий текста. Для английского языка теги соответствуют стандарту Penn Treebank. Теггер корректно определил основные части речи: *[('world', 'NN'), ('become', 'VBP'), ('monstrous', 'JJ'), ('matter', 'NN')]*. Данные частеречного анализа позволяют проводить грамматический анализ, который выявил доминирование в тексте оригинала существительных и глаголов, что типично для повествовательного дискурса. Что касается текста перевода, то почти все слова были помечены как `*NNP*` (имена собственные), например, *[('последнее', 'NNP'), ('время', 'NNP'), ('мир', 'NNP')]*, что свидетельствует о недостаточной адаптации теггера к морфологической структуре русского языка. Применение анализатора pymorphy3 ненамного улучшило картину, поэтому пришлось исправлять ошибки вручную, что не представляется удобным при работе с большим корпусом текстов. В целом, анализ показывает, что разница в распределении частотности частей речи между исходным текстом и его переводом невелика. В текстах оригинала и перевода имена существительные доминируют, занимая первое место по частоте, в то время как глаголы занимают второе место. При дальнейшем исследовании были выявлены случаи грамматических замен в переводе, например, *The world has become* ***monstrous*** *matter-of-fact in latter days. – В последнее время мир стал* ***чудовищно*** *прозаичным* (замена прилагательное на наречие). *A man may properly be hung for* ***murdering****another... – Закон разрешает повесить человека за* ***убийство*** *другого человека…* (замена глагольной формы на существительное), *I* ***felt*** *as if I should have died with vexation. – У меня было* ***ощущение****, что я умру от расстройства* (замена глагола на существительное). Подобного рода замены не повлияли существенно на частотность частей речи в переводе по сравнению с оригиналом.

Среди многочисленных задач квантитативной лингвистики можно выделить составление частотных списков слов, которые формируются, исходя из задач исследования, на основе релевантных текстов или их частей, и отражают частоту обнаруженных элементов. Общим правилом является то, что наиболее часто встречающиеся элементы играют главную роль в тексте и являются значимыми, в то время как элементы с малым количеством упоминаний свидетельствуют об их редком использовании в речи. В целом, «частотные списки позволяют выявить ядро и периферию лексики» [12, с. 741]. Частотный анализ позволяет выявить наиболее употребительные лексемы для определения тематики текста, ключевых персонажей и стилистических предпочтений автора. В художественных текстах частотные имена собственные часто указывают на протагонистов или значимые объекты сюжета. С помощью класса `*FreqDist*` из NLTK в ходе настоящего исследования построены частотные распределения слов и биграмм. Данные демонстрируют, что первые десять самых частотных слов оригинала составляют, в основном, имена собственные: *james, major, grayling, spencer, man, macnab, sparkman, youth, night, mind*. В переводе также доминируют имена: *джеймс, майор, грейлинг, спенсер, мочь, сказать, человек, макнэб, юноша, шотландец*. Одной из причин

расхождений в количестве употребления некоторых лексем в тексте оригинала и перевода является использование в переводе трансформаций, направленных, чаще всего, на сужение значения английского слова, обладающего широким значением. Так, при переводе слова *man* в анализируемом тексте используется девять соответствий: *человек, люди, он, мужчина, тип, незнакомец, кто-нибудь, фермер, постоялец*. Например: *"That's he! That must be the* ***man****!" — Это он! Это наверняка* ***он****! "There was one* ***man*** *stayed with me last night" – Вчера был только один* ***постоялец****…* В ряде случаев переводчик применяет трансформацию опущения, например, *…but Sparkman was a knowing* ***man****… — Но Спаркмен знал, что делает…* Подобная трансформация наблюдается и при передаче глагола движения *go*, переданного в переводе двадцатью соответствиями, среди которых *ехать* встречается 7 раз, *идти* – 6, *отправиться* – 4, *уйти* – 3 , *блуждать* – 2 раза. Остальные соответствия использованы по одному разу. Такое разнообразие переводческих соответствий может быть объяснено тем фактом, что «в русском языке глаголы движения почти всегда выражают информацию не только о движении как таковом, но и о способе движения, в то время как в английском языке информация о способе движения может быть опущена» [13, с. 133]. Например: *Macleod …****hadgone*** *below even while the boat was rapidly approaching the vessel. – Маклеод …* ***спустился*** *вниз, когда увидел, как к пакетботу идет какая-то лодка. … you have* ***tobegoing*** *to England for money… – … какой смысл тебе так далеко, в саму Англию,* ***ехать*** *за деньгами. … I* ***hadgone*** *too far... – … не знал, как далеко* ***ушел***.

В последней части нашего исследования была подсчитана основополагающая количественная характеристика – коэффициент лексического разнообразия (Lexical Diversity Index) — метрика, отражающая соотношение уникальных слов к общему количеству слов в тексте. Высокий коэффициент лексического разнообразия может указывать на богатую лексику, но в контексте сравнения языков он также зависит от морфологической сложности: например, русский язык, будучи флективным, генерирует больше форм одного и того же корня, что искусственно увеличивает количество токенов [14].

Коэффициент лексического разнообразия предполагает «диапазон и вариативность словарного запаса, который говорящий реализует в тексте» [14, p. 459]. Данный показатель позволяет не только охарактеризовать лексическое богатство текста, но также рассмотреть стиль отдельного автора. В основе данного коэффициента лежит соотношение числа уникальных лексических единиц (лемм) и количества их употреблений в тексте (всех словоформ). Лемма – это грамматическая форма, которая используется для представления лексемы [15, p. 611], другими словами, начальная форма слова, например, *go* – это лемма для словоформ *goes, going, went* и т.д. Процедура лемматизации представляет собой довольно сложный процесс, поскольку лемматизаторы допускают погрешности, которые приходится корректировать вручную. В данном исследовании лемматизация русского текста произведена с помощью морфологического анализатора pymorphy3, лемматизация оригинала осуществлена с помощью готовой базы английских лемм WordNetLemmatizer.

Результаты вычисления показали, что коэффициент лексического разнообразия оригинала равен 2.35, коэффициент перевода – 1.67. Меньшее значение коэффициента перевода формально указывает на большее лексическое разнообразие, поскольку коэффициент обратно пропорционален разнообразию (чем выше коэффициент, тем меньше разнообразие). Это наблюдение объясняется как флективной природой русского языка, так и применением в переводе лексической трансформации добавления, что также повышает разнообразие, например: *…one sunny morning in April, their wagon*

started for the city – …в одно прекрасное апрельское утро их фургон двинулся ***по дороге*** *в город. He was always the first to go forward ...– Во всякое* ***дело*** *всегда вызывался первым ... but James, who felt himself equal to any man… – но Джеймс, который чувствовал себя ровным любому* ***взрослому*** *мужчине…*

Еще одной возможностью анализа текста программными средствами является выявление *n*-грамм – последовательности из *n* элементов. *N* -граммы позволяют учесть контекст и основную тему текста. Для выявления *n*-грамм проводится токенизация текста, затем импортируются *n*-граммы. После выполнения этих шагов выводится список *n*-грамм, состоящих, в нашем случае, из двух элементов. Такие *n*-граммы называются биграммами. Каждая биграмма представлена кортежем из двух слов. Самыми часто встречающимися биграммами в оригинале текста оказались биграммы *james grayling, major spencer, joel sparkman, mrs grayling, falmouth packet*. Самые часто встречающиеся биграммы в переводе: *джеймс грейлинг, майор спенсер, миссис грейлинг, майора спенсера, джоуль спаркмен*. Можно заметить, что в русском тексте используется несколько форм одного имени, что объясняется грамматическими особенностями русского языка.

Большое количество в английском оригинале таких биграмм как *sheriff officer, murdered man, ghost story* свидетельствует о детективной и мистической тематике текста. В тексте перевода более частотны другие словосочетания, например, *вполне естественно, потерпел поражение, весь день*, которые не дают определенного представления о теме текста. Данное обстоятельство объясняется тем, что при переводе английских биграмм происходит их трансформация, влекущая за собой изменение грамматической структуры словосочетания. Так, *sheriff officer* переведено как *офицер из ведомства шерифа* (здесь наблюдается перестановка компонентов биграммы и добавление лексических элементов). Другими вариантами перевода этого же словосочетания являются *представитель шерифа* (в данном случае также применена трансформация перестановки) и *о ф и ц е р* (трансформация опущения). При переводе биграммы *murdered man* применяется трансформация опущения – *убитый.* Биграмма *ghost story* переводится с использованием трансформаций перестановки, влекущей за собой добавление предлога: *история о приведениях, история про призрак*, а также трансформации опущения: *As the old lady finished the* ***ghost story****… – Затем почтенная дама закончила свою* ***историю****…* В целом, разница в полученных данных объясняется тем, что «русский язык имеет развитый процесс формирования слов в предложении и словосочетании с помощью формообразующей морфемы в отличие от английского языка, где слова меньше подвержены изменениям, а связь слов в предложении зависит в большей степени от места, а не от формы самого слова» [16, с. 1131]

Заключение

Таким образом, проведенный анализ показал, что инструменты Python и библиотека NLTK являются эффективными средствами для лексико-семантического исследования текстов. Тем не менее, их применимость зависит от языковых особенностей. Английский текст поддается автоматизированному анализу с высокой точностью, в то время как русскоязычные данные требуют дополнительной обработки с использованием морфологических анализаторов, таких как pymorphy3, и ручной коррекции результатов.

Результаты исследования подтверждают гипотезу о том, что перевод, особенно художественный, обладает иной лексико-грамматической структурой по сравнению с оригиналом. Это проявляется в увеличении числа связующих элементов, вариативности форм имен собственных и в определённой степени — в сглаживании тематически

маркированной лексики. Более того, переводческие трансформации (опущение, добавление, перестановка, замена частей речи) систематически влияют на количественные и качественные характеристики текста: лексическое разнообразие возрастает, а ключевые тематические маркеры утрачивают четкость. Данное наблюдение подчеркивает важность комплексного подхода, сочетающего вычислительные методы и лингвистическую интерпретацию, особенно при анализе переводов художественных текстов между языками с различной морфологической структурой.

Библиография

1. Van Der Post H. Natural Language Processing with Python: A comprehensive guide to NLP in the age of AI for 2024. Reactive Publishing, 2023.
2. Hammond M. Python for Linguists. Padstow, Cornwall: TJ International Ltd, 2020.
3. Сафина З.М. Методы квантитативной лингвистики при исследовании оригинала и перевода художественного текста // Филология: научные исследования. 2025. № 10. С. 56-64. DOI: 10.7256/2454-0749.2025.10.76298 EDN: NGJXWK URL: https://nbpublish.com/library_read_article.php?id=76298
4. Rana Y. Python: Simple though an Important Programming language // International Research Journal of Engineering and Technology (IRJET). 2019. Vol. 06, Iss. 2. Pp. 1856–1858.
5. Сафина З. М. Переводческий анализ художественного текста на языке Python // Глобальный научный потенциал. 2024. № 11 (164), Т. 1. С. 177-180. EDN: RTJTGQ.
6. Ладушина М. И. Как язык Python помогает лексикографам // Journal of Applied Linguistics and Lexicography. 2022. Т. 4, № 2. С. 107-121. DOI: 10.33910/2687-0215-2022-4-2-107-121. EDN: UIYYDM.
7. Гагарин С. Н. Базовые методики анализа языковых картин политики с помощью языка программирования Python и библиотеки NLTK (на материалах корпусов британского парламентского дискурса) // Филологические науки в МГИМО. 2024. 10(2). С. 125-140. DOI: 10.24833/2410-2423-2024-2-39-125-140. EDN: GDGMAO.
8. Simms G.W. Grayling; or Murder Will Out // The Wigwam and the Cabin. New York: Redfield, 1856. Pp. 2-36.
9. Симмс У. Г. Грейлинг, или "Тайное становится явным" (пер. М. Л. Павлычевой) // Вигвам и хижина. Санкт-Петербург: Дмитрий Буланин, 2018. С. 27-55.
10. Bird S., Klein E., Loper E. Natural Language Processing with Python. O'Reilly Media, 2009.
11. Хайрова Н. Ф., Мамырбаев О. Ж., Петрасова С. В., Мухсина К. Ж. Современные технологии обработки текстовых данных на базе пакета NLTK Python: учеб. пособ. Харьков: ООО "В деле", 2020.
12. Сафина З. М., Корнилова А. Д., Смакова А. Л. Количественный и статистический анализ лексических единиц в художественном переводе // Вестник Башкирского университета. 2022. Т. 27, № 3. С. 741-746. DOI: 10.33184/bulletin-bsu-2022.3.42. EDN: FGZGYW.
13. Морозкина Е. А., Воробьев В. В., Сафина З. М. Статистические методы исследования в художественном переводе // Доклады Башкирского университета. 2023. Т. 8, № 3. С. 130-137. DOI: 10.33184/dokbsu-2023.3.15. EDN: KHORTY.
14. McCarthy P.M., Jarvis S. vocd: A theoretical and empirical evaluation // Language Testing. 2007. 24 (4). Pp. 459-488. DOI: 10.1177/0265532207080767. EDN: JTJVXN.
15. Jurafsky D., Martin J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey: Prentice Hall, 2009.

16. Мифтахова Р. Г., Морозкина Е. А. Нейронное представление семантического поля // Вестник Башкирского университета. 2021. Т. 26, № 4. С. 1130–1135. DOI: 10.33184/bulletin-bsu-2021.4.48. EDN: KWAPJJ.

Результаты процедуры рецензирования статьи

Рецензия выполнена специалистами Национального Института Научного Рецензирования по заказу ООО "НБ-Медиа".
В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов можно ознакомиться здесь.

В рецензируемой статье анализируются возможности библиотеки Natural Language Toolkit (NLTK, версия 3.8) на языке программирования Python 3.11 при проведении лексико-семантического анализа художественного текста и его перевода. Актуальность работы не вызывает сомнения. Как верно отмечается, «обработка естественного языка является одной из высокоприоритетных задач современной лингвистики, что во многом связано с необходимостью получения быстрых и качественных результатов анализа большого массива данных», «языки программирования позволяют значительно ускорить работу лингвистов, систематизировать результаты сбора данных в упорядоченном виде, а также, в отличие от традиционных методов, могут гарантировать точность и объективность полученных результатов».
Теоретической основой обоснованно послужили работы отечественных и зарубежных ученых по компьютерной лингвистике, вопросам применения методов обработки естественного языка и использованию языка программирования Python в лингвистике. Библиография насчитывает 12 источников, в том числе художественные, соответствует специфике изучаемого предмета, содержательным требованиям и находит отражение на страницах статьи. Однако теоретическая часть, так называемый обзор литературы, по мнению рецензента, раскрыта недостаточно. Также следует отметить, что несмотря на высокую актуальность изучаемой проблематики, автор(ы), апеллируют к узкому кругу работ российских авторов (все с участием З. М. Сафиной). Рекомендуем проанализировать исследования в данной области и других отечественных ученых.
Методология исследования продиктована поставленной целью и носит комплексный характер: использованы общенаучные методы анализа и синтеза, описательный метод, включающий наблюдение, обобщение, интерпретацию, классификацию, сравнительно-сопоставительный и статистический методы, а также методы обработки естественного языка. Материалом исследования послужило произведение американского писателя XIX века Уильяма Гилмора Симмса «Grayling: or, "Murder Will Out"»(1841) и его русский перевод «Грейлинг, или "Тайное станет явным"», выполненный М. Л. Павлычевой).
В ходе анализа рассмотрены особенности лексико-семантического анализа художественного текста с использованием библиотеки NLTK на языке Python: нормализация текста, частеречное тегирование, частотный анализ, коэффициент лексического разнообразия, выявление n-грамм. Полученные результаты подтвердили гипотезу о том, что перевод, особенно художественный, обладает иной лексико-грамматической структурой по сравнению с оригиналом, причем «переводческие трансформации систематически влияют на количественные и качественные характеристики текста: лексическое разнообразие возрастает, а ключевые тематические маркеры утрачивают четкость». В заключении отмечено, что, несмотря на эффективность инструментов Python и библиотеки NLTK в лексико-семантическом исследовании художественного текста, «их применимость зависит от языковых особенностей»: «английский текст поддается автоматизированному анализу с высокой точностью, в то

время как русскоязычные данные требуют дополнительной обработки с использованием морфологических анализаторов и ручной коррекции результатов».

Теоретическая значимость исследования связана с определенным вкладом результатов проделанной работы в развитие таких современных научных направлений, как когнитивная лингвистика, контрастивная лингвистика, корпусная лингвистика; методологии квантитативной лингвистики. Практическая значимость заключается в возможности использования ее результатов при проведении автоматизированного лексико-семантического исследования художественного текста и его перевода.

Статья имеет четкую, логически выстроенную структуру (введение, основная часть, заключение). Тем не менее, объем рукописи близок к минимальным требованиям редакции. Также, по мнению рецензента, название статьи требует уточнения: в представленном материале речь о лексико-семантическом анализе художественного текста и его переводе с использованием инструментов Python.

Стиль изложения отвечает требованиям научного описания. Рукопись имеет завершенный вид; она вполне самостоятельна, оригинальна, будет интересна и полезна широкому кругу лиц и может быть рекомендована к публикации в научном журнале «Филология: научные исследования» после внесения соответствующих правок.

Результаты процедуры повторного рецензирования статьи

Рецензия выполнена специалистами Национального Института Научного Рецензирования по заказу ООО "НБ-Медиа".

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.

Со списком рецензентов можно ознакомиться здесь.

Рецензируемая статья посвящена лексико-семантическому анализу оригинала и перевода художественного текста с применением инструментов Python и библиотеки NLTK, что позволяет отнести её к междисциплинарным исследованиям на стыке лингвистики, цифровых гуманитарных наук и переводоведения. Предмет исследования сформулирован чётко и последовательно: автор ставит цель продемонстрировать возможности инструментов NLP (Natural Language Processing) для анализа художественного текста и его перевода, а также выявить переводческие трансформации и их влияние на количественные и качественные параметры текста. Работа опирается на сопоставление оригинального произведения У. Г. Симмса и его русского перевода, что делает исследование методологически выверенным и релевантным задачам современного компьютерного текстового анализа.

Методология исследования отличается комплексностью и системностью. Автор последовательно описывает этапы обработки текста: нормализацию, токенизацию, удаление пунктуации, лемматизацию, исключение стоп-слов, частеречное тегирование, построение частотных списков, анализ биграмм и вычисление коэффициента лексического разнообразия. Подробное изложение каждого этапа делает методику воспроизводимой и научно достоверной. Особое внимание уделено трудностям обработки русскоязычных текстов: показано, что стандартные инструменты NLTK адекватно работают с английскими данными, но дают значительные ошибки при анализе русского материала, требуя дополнительного использования морфологических анализаторов (pymorphy3) и ручной корректировки тегов. Это наблюдение важно для всей области русскоязычного NLP и корректно обозначено автором как

методологическое ограничение исследования.

Актуальность работы определяется сразу несколькими факторами: растущей ролью цифровых методов в лингвистике, интересом к количественным параметрам художественного текста, потребностью в объективных критериях анализа перевода и очевидной недостаточной разработанностью этой темы в отечественной филологии. Автор обоснованно отмечает, что исследования, сочетающие компьютерный анализ и художественный перевод, в российской и зарубежной науке пока немногочисленны, а потому предложенная работа заполняет важный исследовательский пробел.

Научная новизна статьи проявляется в сочетании нескольких подходов: сопоставлении частеречной характеристики оригинала и перевода, анализе частотности слов и биграмм, выявлении корреляции между переводческими трансформациями и изменением лексического разнообразия текста. Автор демонстрирует, что русскоязычный перевод использует значительно более вариативный словарь: для передачи английского man задействовано девять лексических соответствий, глагол go передаётся более чем 20 эквивалентами. Такой подход позволил сделать важный вывод: перевод не только меняет структуру текста, но и формирует иной семантический рельеф произведения. Это существенный вклад в переводоведение и компьютерную лингвистику.

Стиль и структура статьи отличаются логичностью, научной чёткостью и последовательностью. Текст разбит на тематические блоки, плавно переходящие от теоретической части к аналитической. Описание инструментов и этапов анализа выполнено корректно, без излишней технической детализации, что делает статью доступной для широкого круга исследователей. Большим достоинством является наличие конкретных примеров из текстов оригинала и перевода — они наглядно иллюстрируют выявленные закономерности. В то же время можно рекомендовать более явно выделить промежуточные аналитические выводы после каждого крупного этапа исследования, чтобы усилить структурную целостность текста.

Библиография включает широкий спектр источников: труды по NLP, переводоведению, квантитативной лингвистике, современные статьи, учебные пособия и программную документацию. Такой подбор литературы подтверждает междисциплинарный характер работы и демонстрирует её опору на актуальные научные исследования. Источники корректно отражают теоретическую базу исследования и подтверждают авторскую интерпретацию данных.

Апелляция к оппонентам проявляется в обсуждении современных подходов к компьютерному анализу текста и критическом рассмотрении ограничений существующих инструментов. Автор корректно указывает на недостатки моделей POS-тегирования для русского языка и предлагает пути их компенсации, что свидетельствует о профессиональном владении предметом исследования.

Выводы статьи обоснованы и логически следуют из проведённого анализа. Автор показывает, что применение Python и NLTK позволяет выявить ключевые закономерности перевода, отличия в распределении частей речи, структуре биграмм, частотности слов, а также вычислить коэффициент лексического разнообразия, который демонстрирует отличия между переводом и оригиналом. Исследование представляет несомненный интерес для лингвистов, переводоведов, специалистов по цифровым гуманитарным исследованиям и разработчиков NLP-инструментов.

Работа является зрелым, качественным и актуальным исследованием, и может быть рекомендована к публикации в научном журнале.