

DOI: 10.33184/dokbsu-2025.2.6

Статистическое исследование текста программными средствами (на материале оригинала и перевода новеллы Э. А. По «Падение дома Ашеро́в»)

Е. А. Морозкина*, Е. П. Хабибуллина

Уфимский университет науки и технологий

Россия, Республика Башкортостан, 450076 г. Уфа, ул. Заки Валиди, 32.

**Email: morozkinaea@mail.ru*

В работе проведено статистическое исследование оригинала и перевода новеллы Э. А. По «Падение дома Ашеро́в» с использованием программы, написанной на языке программирования «Python». В исследовании используются такие методы обработки естественного языка, как токенизация, лемматизация, что позволяет определить лексическое разнообразие и информативность текстов, а также выявить ключевые лексемы в текстах оригинала и перевода, определить полярность текста.

Ключевые слова: статистический метод, перевод художественного текста, сравнительный анализ, ключевые слова, полярность.

В последние годы наблюдается повышенный интерес ученых к применению прикладных методов компьютерной лингвистики для анализа частотности употребления лексем в художественных текстах, что позволяет выявить такие критерии, как лексическое разнообразие и информативность. Определение таких критериев текста обусловило актуальность темы исследования и обеспечило новизну полученных выводов.

Цель исследования – произвести сравнительный анализ оригинала и перевода художественного текста на русский язык с использованием статистических методов на базе языка программирования Python. Задачи исследования включают проведение предварительной обработки текста методами токенизации и лемматизации и далее осуществление подсчета необходимых параметров.

В качестве материала исследования использована новелла Э. А. По «Падение дома Ашеро́в» («The Fall of the House of Usher»), опубликованная в 1839 г., а также ее перевод на русский язык, выполненный Н. Галь и опубликованный в 1972 г.

Как отмечается современными исследователями, «на сегодняшний день инструменты математической статистики причисляются ...» к методам «современного

лингвистического анализа текста» [3, с. 130]. Для того чтобы программа подвергла обработке естественный язык, необходимо произвести предварительную работу над текстом, что и было осуществлено при помощи инструментов языка программирования «Python». Как справедливо отмечалось Е. А. Морозкиной, В. В. Воробьевым, А. Д. Корниловой, «современные технологии позволяют быстро и автоматизировано провести предварительную обработку художественного текста» [4, с. 10]. В первую очередь тексты оригинала и перевода были переведены в текстовые файлы в формате .txt с кодировкой Unicode (UTF-8). Далее каждый текст был разбит на токены по принципу: одно слово или один знак препинания – один токен. Далее из текста были исключены верхний регистр, знаки препинания и так называемые «стоп-слова». Отметим, что в каждом языке существует свой список таких стоп-слов в силу различия грамматических структур языков. Например, “и”, “на” и другие в русском; “a”, “the”, “and” и некоторые другие в английском языке. Кроме того, была произведена лемматизация текста, в результате чего разные словоформы одного и того же слова не учитываются программой как разные слова и будут восприняты при анализе как одна лексема. Наконец, был произведен подсчет таких параметров анализа текста, как коэффициент лексического разнообразия, процент информативности в оригинале и переводе анализируемого художественного текста. Кроме того, были определены ключевые слова художественного текста, под которыми мы понимаем слова, обладающие высокой частотностью и важными смысловыми характеристиками.

В большинстве способов измерения при расчете коэффициента лексического разнообразия учитывается соотношение количества всех словоформ и количества уникальных лексических единиц. Именно такой подход был использован в настоящем исследовании. Высокое лексическое разнообразие означает, что автор использует множество различных слов, что делает его текст более интересным и выразительным. Подсчет информативности текста предполагает поиск соотношения уникальных («полезных») слов к общему количеству слов в тексте [5, с. 18]. Информативность позволяет оценить, насколько существенную информацию содержит текст относительно общего объема лексических единиц.

Сравнительный анализ текстов оригинала и перевода, проведенный с помощью программы, разработанной на языке программирования Python, позволяет выявить ключевые слова, которые в зависимости от контекста использования могут быть интерпретированы по-разному. Например, в работе Майкла Стаббса «Три концепта ключевых слов» (“Three Concepts of Keywords”) описаны три концепта ключевых слов, а именно: ключевые слова в контексте культурологии, которые имеют культурное значение и могут помочь в создании концептуальных категорий; ключевые слова в корпусной лингвистике, которые предполагают статистически значимые слова в определенных текстах; и наконец ключевые слова, которые являются частью

общепринятых фраз и выражают общие культурные ценности [6, с. 4]. В рамках статистического исследования программными средствами двух текстов оригинала и перевода под ключевыми словами, как отмечалось, понимаются наиболее частотные слова, использованные в тексте, а также слова, содержащие определяющие для данного текста смысловые значения.

Подсчет лексического разнообразия, информативности, а также нахождение ключевых слов были осуществлены с помощью разработанной нами версии программы, написанной на языке «Python». При этом предварительно программа была проверена на текстах небольшого объема для того, чтобы убедиться в достоверности результатов обработки текстов с помощью созданной компьютерной программы. Результаты анализа оригинала и перевода исследуемого художественного текста представлены в таблице 1.

Таблица

Сравнительный анализ компонентов оригинала и перевода новеллы Э. А. По «Падение дома Ашер»

Параметр подсчета	Оригинал Э. А. По	Перевод Н. Галь
1. Лексическое разнообразие	3.8767	2.7973
2. Информативность	0.2430	0.3435
	<i>Usher</i> – 22	<i>Ашер</i> – 37
3. Ключевые слова (самые частотные и важные по смыслу слова)	<i>long</i> – 20	<i>дом</i> – 24
	<i>one</i> – 15	<i>стена</i> – 15
	<i>house</i> – 14	<i>мочь</i> – 15

В табл. коэффициент лексического разнообразия текста оригинала составляет 3.8767, текста перевода – 2.7973; информативность текста оригинала составляет 0.2430, перевода – 0.3435. Также представлены ключевые слова. Показательно, что не все ключевые лексемы оригинала совпали с ключевыми лексемами перевода. В частности, в оригинале наиболее частотными оказались: *Usher*, *long*, *one*, *house*, а в переводе: *Ашер*, *дом*, *стена*, *мочь*. Из таблицы видно, что коэффициент лексического разнообразия в переводе оказался ниже, чем в тексте оригинала, что говорит о более высоком лексическом разнообразии текста перевода. Соответственно, информативность текста перевода оказалась выше, чем информативность текста оригинала. Означенные количественные данные, полученные эмпирическим путем с использованием языка программирования «Python», подтверждают тот факт, что лексическое разнообразие того или иного языка в определенном отношении обусловлено спецификой структуры языка, что, безусловно, проявляется в отношении английского и русского языков, которые, как известно, являются разноструктурными.

При анализе ключевых слов наибольший интерес представляет выявленный нами фактор количественного разрыва в использовании имени собственного *Usher* – 22 раза в оригинале и 37 раз в переводе. Такое несоответствие, полагаем, обусловлено несколькими факторами: во-первых, зачастую в английском оригинале вместо имени собственного были использованы личные местоимения, которые в некоторых случаях в переводе переводчик заменил на имя собственное. Например, *He admitted...* / ...**Ашер** признался; *he had been sitting...* / ...**Ашер** сидел...; *It was thus that he spoke of the object...* / Именно так говорил **Родерик Ашер** о...; во-вторых, при переводе некоторых конструкций переводчик добавлял имя собственное *Ашер* там, где в английском оригинале находилось другое существительное или словосочетание: *...his very ancient family...* / ...древний и что все **Ашеры**...; *The writer spoke of...* / **Ашер** писал...; *...physician of the family...* / ...врач **Ашеров**; *...constitutional and a family evil...* / ...проклятие ... семьи ... **Ашеров**. Таким образом, переводчик употребил имя собственное *Ашер* значительно большее количество раз, чем это слово было использовано в оригинале, что связано с различающейся структурой языков, а также с тем, что *Ашер* – это ключевой персонаж в произведении, и поэтому переводчик, возможно, таким образом хотел ярче передать психологическое состояние героя. Кроме того, переводчик счел необходимым ввести пояснения к лексемам *family* (семья), *writer* (писатель), добавив к ним имя собственное *Ашер*.

Определенный интерес для анализа представляет лексема *house*, так как она является ключевой как в оригинале, так и в переводе художественного произведения. Лексема *house* встретилась 14 раз в тексте оригинала, а ее соответствие в переводе *дом* встретилось 24 раза. Для того чтобы раскрыть причины такого частотного расхождения, обратимся к словарям и сравним лексические значения лексемной пары *house-дом*. В словарной статье, посвященной лексеме *house* [7], даны семь вариантов значений, из которых нами были выбраны значения под цифрами 1 (a, b) и 7, поскольку именно в этих значениях была использована лексема *house* в исследуемом художественном тексте: «house – 1 a [count]: a building in which a family lives. b [singular]: the people who live in a house... 7 [count]: a royal or noble family including ancestors and all the people who are related to them». В Большом толковом словаре русского языка под редакцией С. А. Кузнецова [8] находим пять типов значений лексемы *дом*, среди которых были выбраны первые четыре значения, поскольку именно они соответствуют значениям, которые были использованы в тексте перевода новеллы на русский язык: «1. Здание, строение, предназначенное для жилья, для размещения различных учреждений и предприятий. 2. Жилое помещение, квартира; жилье. 3. Семья, люди, живущие вместе, одним хозяйством. 4. Устар. Род, фамилия; поколение». Как видно из сопоставления словарных статей, двум значениям английского слова *house* соответствует четыре словарных значения слова *дом*. Таким образом, употребление русским переводчиком слова *дом* обусловлено большей вариативностью словарных значений лексемы *дом* в русском языке.

Кроме того, значительное расхождение в количестве употреблений ключевой лексики *house* и его словарного соответствия *дом* происходит в силу того, что словом *дом* в русском тексте переводятся также и английские лексемы: *mansion* (7), *dwelling* (3), *building* (6), *estate* (2), *masonry* (1) – в скобках отмечено количество наличия в тексте означенных лексем – хотя при этом переводчик также использует и другие русские эквиваленты: *обитель* (2), *имение* (1), *поместье* (1), *жилище* (2), хотя, как можно заметить, их употребление оказывается менее частотным: *that lay in unruffled lustre by the dwelling...* / ...недвижная гладь едва поблескивала возле самого *дома*; *in this mansion of gloom...* / ...в этой *обители* уныния...; *about the whole mansion and domain...* / ...над этим *домом*, усадьбой; *...the real aspect of the building...* / ...в подлинный облик *дома*; *No portion of the masonry had fallen* / Однако нельзя было сказать, что *дом* совсем пришел в упадок; *the fabric gave little token of instability...* / ...снаружи вовсе нельзя было заподозрить, будто *дом* непрочен. Как видим, синонимии в оригинале соответствует в переводе на русский язык преимущественно эквивалент *дом*. Лексема *дом* является ключевой и смыслообразующей в исследуемом художественном тексте. Вместе с тем, как видим, лексемы, входящие в синонимический ряд лексемы *house/дом* могут до некоторой степени «менять свое значение в зависимости от контекста» [9, с. 70]. Лексема представляется настолько важной, что она вынесена в заглавие текста. Действительно, *дом* является важной составляющей жизни любого человека, т.к. место, где человек родился и вырос, оказывает огромное влияние на развитие личности. *Дом* символизирует тепло семейных отношений, поддержку близких, что предполагает влияние на психологическое состояние человека. *Дом* способствует формированию у человека чувства безопасности, опоры, ощущения принадлежности фамилии, клану. Употребив лексему *дом* при переводе таких английских существительных, как *house, dwelling, mansion, masonry*, переводчик, вероятно, предполагал обратить особое внимание реципиента на *дом* как наиболее важный смысловой компонент.

Таким образом, статистическое исследование текста оригинала и перевода новеллы Э. А. По «Падение дома Ашеров» с помощью компьютерной программы, написанной на языке Python, показало, что лексическое разнообразие текста перевода оказалось существенно выше текста оригинала, в результате чего можно утверждать, что русскоязычный перевод лексически богаче, чем текст оригинала, что может быть связано с разноструктурностью исследуемых языков. Вполне закономерно, что и информативность текста перевода оказывается несколько выше информативности оригинала. Кроме того, была выявлена существенная разница в употреблении языковых лексем в качестве ключевых, многие из которых не совпали по частотности использования в текстах оригинала и перевода. Данное исследование позволило также определить, что ключевая лексема *дом* в тексте перевода оказалась значительно более частотной, чем лексема *house* оригинале, что в определенном отношении может быть

связано как с лингвистическими особенностями теста, так и со спецификой менталитета представителей разных культур.

Литература

1. Poe E. A. The fall of the House Usher. URL: <https://www.ibiblio.org/ebooks/Poe/Usher.pdf>
2. По Э. А. Падение дома Ашеров. URL: <https://lib.ru/INOFAANT/POE/asher.txt>
3. 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.
4. Морозкина Е. А., Воробьев В. В., Корнилова А. Д. Корпусное исследование семантики глагола в структуре ремарок драматургического текста (на материале оригинала и англоязычного перевода пьесы А.П. Чехова «Вишневый сад») // Доклады Башкирского университета. 2024. Т. 9. №4. С. 8–15.
5. Васильев Ю. Обработка естественного языка. Python и spaCy на практике. СПб.: Питер, 2021. С. 256.
6. Stubbs M. Three concepts of keywords / In M. Bondi & M. Scott (eds) *Keyness in Texts*. Amsterdam: Benjamins, 2010. P. 21–42.
7. The Britannica Dictionary. URL: <https://www.britannica.com/dictionary/house>
8. Кузнецов С. А. Большой толковый словарь русского языка. URL: <https://gramota.ru/biblioteka/slovari/bolshoj-tolkovyj-slovar>
9. Морозкина Е. А., Воробьев В. В., Щанькин А. П. Лексико-семантическое после «Финансы» в оригинале и переводе романа Т. Драйзера «Финансист» // Доклады Башкирского университета. 2025. Т. 10. №1. С. 67–72.

Статья рекомендована к печати кафедрой
лингводидактики и переводоведения
Уфимского университета науки и технологий
(д-р филол, наук, проф. Е. А. Морозкина).

Statistical study of the text by software (based on the original and translation of E. A. Poe's novella “The Fall of the House of Usher”)

E. A. Morozkina*, E. P. Khabibullina

Ufa University of Science and Technology
32 Zaki Validi st., 450076 Ufa, Republic of Bashkortostan, Russia.

**Email: morozkinaea@mail.ru*

The article presents a statistical study of the original and translation of E. A. Poe's novella “The Fall of the House of Usher” with the use of a program written in the Python programming language. The research uses natural language processing methods such as tokenization and lemmatization, which allows us to determine the lexical diversity and informative value of texts, as well as identify key lexemes in the original and translated texts and determine the polarity of the text.

Keywords: statistical method, translation of a literary text, comparative analysis, keywords, polarity.