DOI: 10.33184/dokbsu-2023.2.8

Автоматическая обработка художественных текстов А. П. Чехова и их англоязычных переводов с опорой на методы лемматизации и частеречной разметки

Е. А. Морозкина, А. Д. Корнилова*

Уфимский университет науки и технологий
Россия, Республика Башкортостан, 450076 г. Уфа, ул. Заки Валиди,32

**Email: anastasia.ufa@mail.ru*

В статье предлагается использовать методы автоматической обработки текстов для проведения сопоставительного анализа оригиналов художественных произведений и их переводов с целью выявления коэффициента лексического разнообразия. Установлено, что пьеса А. П. Чехова «Вишневый сад» и ряд его рассказов лексически вариативнее их англоязычных версий перевода. Выяснилось, что в пьесе А. П. Чехова «Вишневый сад», в отличие от рассказов, глагольная лексика употребляется чаще других знаменательных частей речи. Предпринята попытка представления зависимости коэффициента лексического разнообразия от объема текста в виде линейной регрессии.

Ключевые слова: лемматизация, частеречная разметка, автоматическая обработка текстов.

С появлением интернета и компьютерных программ стало возможным проведение анализа переводов с точки зрения их близости к оригиналу методами автоматической обработки текстов, которые относятся к области компьютерной лингвистики. Анализ текстов проводится на лексико-морфологическом уровне с помощью методов лемматизации и частеречной разметки с целью получения лемм заданных словоформ и их грамматических признаков. Под лемматизацией подразумевается процесс образования первоначальной формы слова (леммы) для словоформ текста. Например, формы глаголов *go*, *goes*, *went*, *gone*, *going* сводятся к одной лемме *go*.Метод лемматизации упрощает дальнейшую работу исследователя, поскольку он позволяет выявить все ассоциации с исследуемым словом [1, с. 1131], т.е. все словоформы одной конкретной леммы, что особенно актуально для флективных языков, к которым относятся русский и английский языки. Частеречная разметка рассматривается как более трудоемкий процесс, в котором каждой лексической единице приписываются определенные грамматические характеристики, включая часть речи, лемму и набор граммем (грамматических категорий). Лемматизация и частеречная разметка широко применяются при составлении корпусов текстов, с помощью которых можно провести статистический анализ художественных произведений и выявить их лингвистические особенности. Более того, корпуса текстов, в частности параллельные корпуса, могут быть использованы при обучении, к примеру, нейронной сети, которая анализирует параллельные корпуса с целью обнаружения в них определенных закономерностей [2, с. 500].

В рамках данной статьи проводится сопоставительный анализ художественных текстов методами лемматизации и частеречной разметки в программах MyStem, AntConc, TagAnt. Первая программа, разработанная компанией «Яндекс», осуществляет морфологический разбор русских слов [3]. Две другие программы являются авторской разработкой профессора университета Васэда Э. Лоуренса, где AntConc – это мультиплатформенный инструмент для проведения корпусных лингвистических исследований, TagAnt – это многоязычный сегментный теггер частей речи [4]. Материалом исследования послужили пьеса А. П. Чехова «Вишневый сад» (1904) [5] и два ее перевода на английский язык, осуществленные британской переводчицей К. Гарнетт (1911) [6] и группой американских переводчиков Р. Нельсоном, Р. Пивером и Л. Волохонской (2015) [7], а также ряд рассказов А. П. Чехова [5] и их переводы на английский язык, осуществленные К. Гарнетт [8], Р. Пивером и Л. Волохонской [9].Специально выделены три рассказа А. П. Чехова «Скучная история», «Палата №6», «В овраге», перевод которых был осуществлен К. Гарнетт, Р. Пивером и Л. Волохонской. Рассказы «Скучная история», «Палата №6», «В овраге» выбраны для анализа в связи с тем, что по общему количеству лексических единиц эти рассказы примерно соответствуют объему пьесы А. П. Чехова «Вишневый сад», что обеспечивает наглядность графика линейной регрессии. Остальные двадцать семь рассказов выбраны методом случайной выборки. На первом этапе исследования к оригиналам и переводам применяется метод лемматизации в программах MyStem (для оригиналов) и AntConc (для переводов). На втором этапе из списка лемм посредством собственноручно составленного списка стоп-слов исключаются служебные части речи и некоторые имена собственные. На третьем этапе к полученным леммам применяется метод частеречной разметки, в результате которого к леммам присоединяются соответствующие части речи и набор граммем в программах MyStem (для оригиналов) и TagAnt (для переводов). После последовательного выполнения методов лемматизации и частеречной разметки получается готовый частотный список лемм, разбитый на знаменательные части речи.

Как правило, основной целью сопоставительного исследования языков является «установление общего и различного в языках сравнения, в то время как дальняя (стратегическая) цель состоит в выявлении языковых универсалий» [10, с. 441]. Как замечает Н. П. Пешкова: «Существует мнение, что преимуществом количественных методов являются, прежде всего, их надежность и достоверность, способствующие объективности» [11, с. 317]. И, действительно, полученный частотный список позволяет предварительно сопоставить и оценить частотность лемм оригинала и переводов, что помогает заметить как лексические, так и грамматические отличительные черты текстов. В табл. 1 представлено сравнение количественной составляющей знаменательных частей речи оригиналов и двух версий переводов пьесы А. П. Чехова «Вишневый сад» и его рассказов «Скучная история», «Палата №6», «В овраге».

Таблица 1. Количественная составляющая частей речи
оригиналов текстов А. П. Чехова и их англоязычных переводов

Тексты	**V**	**N**	**ADV**	**A**	**Итого**
А. П. Чехов					
Вишневый сад	2535	2262	764	661	6222
Скучная история	3086	3736	1015	1310	9147
Палата №6	2851	3521	853	1221	8446
В овраге	2316	2526	637	717	6196
К. Гарнетт					
The Cherry Orchard	4280	2492	696	820	8288
Dreary Story	5076	4162	848	1417	11503
Ward №6	4256	3752	732	1152	9892
In the Ravine	3515	2829	601	897	7842
*Р. Нельсон, Р. Пивер и Л. Волохонская Р. Пивер и Л. Волохонская					
The Cherry Orchard*	3970	2426	675	820	7891
Boring Story	4485	3890	831	1372	10578
Ward no.6	3834	3746	772	1148	9500
In the Ravine	3160	2778	556	869	7363

Как можно заметить, в оригинале пьесы «Вишневый сад» глагольная лексика (V) употребляется чаще существительных (N), а наречия (ADV) – чаще прилагательных (A), в то время как в трех рассказах А. П. Чехова наблюдается иная тенденция – существительные употребляются чаще глаголов, а прилагательные– чаще наречий. В обоих англоязычных версиях пьеса “TheCherryOrchard” содержит больше глаголов, чем существительных, и больше прилагательных, чем наречий, тогда как в переводных рассказах чаще употребляются глаголы, затем следуют существительные, прилагательные и наречия. Отличия в частотности употребления тех или иных знаменательных частей речи наталкивает на предположение, что в пьесе А. П. Чехова глаголы играют особенно важную роль, в отличие от его рассказов. Вероятно, это связано с особенностью структуры драматургического произведения, в котором центральное положение занимают диалоги, монологи и ремарки (специальные указатели). В последних зачастую при помощи глаголов описывается действие персонажа. Также, сравнивая оригиналы и переводы пьесы «Вишневый сад» и трех рассказов, можно заметить, что в оригиналах присутствует меньше знаменательных частей речи, чем в версиях перевода, причем суще-

ственное количественное расхождение можно наблюдать на примере частотности глаголов, которые чаще встречаются у К. Гарнетт, затем у группы американских переводчиков и только потом у А. П. Чехова. Высокая частотность употребления английских глаголов может быть связана с особенностью строя английского языка, в котором для выражения некоторых комплексных грамматических категорий используются вспомогательные глаголы *be* и *have*. В оригинале пьесы глагол *быть* употребляется 120 раз, глагол *иметь* – 4 раза, тогда как в британском тексте их эквиваленты встречаются 928 и 278 раз соответственно, а в американском – 841 и 233 раза. Например: *Душечкамоя* ***приехала*** *[5]! My little darling* ***has come*** *back [6]! My darling* ***has come****[7].*Кроме того, высокая частотность глаголов *be* и *have* может быть объяснена тем фактом, что данные глаголы нередко употребляются в качестве глагола связки составного именного сказуемого, например: *Она* ***хорошая, добрая, славная****... [5].She* ***is good****, and* ***kind****, and* ***nice****... [6]. She's* ***good, kind, nice****... [7].* Также встречаются случаи, когда глагол *have* используется в тексте в качестве модальных глаголов *have to* и *have got to*, например: *Мне сейчас ... в Харьков ехать[5]. I* ***have to*** *set off for Harkov... [6]. I'****ve got to*** *leave for Kharkov... [7].* Стоит отметить, что в британском переводе вспомогательные глаголы *be* и *have* употребляются чаще, чем в американском переводе, что может быть связано с тем, что «американские переводчики чаще склоняются к использованию простых грамматических конструкций» [12, с. 127]. Переводчикам важно «опираться на широкий кругозор, хорошее знание художественных текстов, принадлежащих данному автору», чтобы избрать верную стратегию перевода [13, с. 428]. Таким образом, русский и английский языки имеют «разные структурные способы словообразования и иные установки в аспекте синтаксических связей и порядка слов» [14, с. 779].

Чрезвычайно важно также определить степень лексического разнообразия оригинала и его переводов. Тема лексического разнообразия широко рассматривается в работах Л. Л. Гонсалвес, Л. Б. Гонсалвес, Д. Грива, Ф. М. Маккарти, С. Джарвиса, Д. М. Бузаджи и В. К. Ланчикова[15–18]. Определение коэффициента лексического разнообразия необходимо, чтобывыявить«степень богатства словаря автора и его вариативности» [19, с. 37]. Согласно исследователям Л. Л. Гонсалвес и Л. Б. Гонсалвес, степень лексического разнообразия можно посчитать по формуле

$$k = 100 * \frac{n}{N},$$

где N – общее количество всех словоупотреблений корпуса, n – количество уникальных слов в корпусе (т.е. таких слов, повторы которых не учитываются), k – коэффициент лексического разнообразия [15, с. 558]. В нашем исследовании в качестве показателя N представлено общее количество всех знаменательных и служебных частей речи, тогда как в качестве показателя n – только уникальные знаменательные части речи. В случае показателя n важно учитывать знаменательные части речи, поскольку во время прочтения художественного текста именно они дают главную информацию, тогда как служебные части речи помогают ее связать, но особой информативной ценности они не несут.

В табл. 2 представлены результаты степени лексического разнообразия оригиналов и англоязычных переводов пьесы А. П. Чехова «Вишневый сад», а также его рассказов «Скучная история», «Палата №6», «В овраге».

Таблица 2. Лексическое разнообразие оригиналов и
их англоязычных переводов художественных текстов А. П. Чехова

Тексты	**N**	**n**	**k**
А. П. Чехов			
Вишневый сад	13103	1964	14.99
Скучная история	18128	3301	18.21
Палата №6	16097	3162	19.64
В овраге	11939	2299	19.26
К. Гарнетт			
The Cherry Orchard	18298	1633	8.92
Dreary Story	24232	2871	11.85
Ward №6	21569	2687	12.46
In the Ravine	16416	1986	12.10
*Р. Нельсон, Р. Пивер и Л. Волохонская Р. Пивер и Л. Волохонская			
The Cherry Orchard*	17501	1554	8.88
Boring Story	22238	2828	12.72
Ward no.6	19891	2665	13.40
In the Ravine	15222	1933	12.70

При сравнении полученных результатов становится понятно, что «Палата №6» лексически разнообразнее других текстов А. П. Чехова. То же самое отмечается и в переводах, где “Ward №6” и “Ward no.6” лексически вариативнее других переводных текстов. Это говорит о том, что чем выше показатель *k*, тем богаче наполнен текст разнообразной лексикой. Стоит отметить, что на показатель *k* влияет объем текста, т.е. чем он объемнее, тем больше в нем содержится повторов одних и тех же лемм [20, с. 743] и, соответственно, тем ниже коэффициент лексического разнообразия. Представленные результаты демонстрируют одну интересную особенность: оригинал и переводы пьесы «Вишневый сад» по степени лексического разнообразия (*k*) уступают трем другим рассказам, причем по объему текста (*N*) «Вишневый сад» и три рассказа примерно схожи. В оригинале и двух переводах пьесы А. П. Чехова содержится меньшее количество уникальных лексических единиц (*n*) по сравнению с рассказами. Подобное явление можно

объяснить тем, что текст пьес менее лексически вариативен, поскольку в нем фигурируют диалоги, тогда как описание практически отсутствует. В диалогах также присутствуют выразительные лексические средства, но их сравнительно меньше, чем в описаниях, поскольку в репликах персонажей чаще заметны повторы одних и тех же слов, также устная форма общения богата именно единообразной разговорной, а не вариативной литературной лексикой. Любопытно и то, что британская версия “The Cherry Orchard” лексически богаче американской, тогда как в переводных рассказах наблюдается иная тенденция – американские версии “Boring Story”, “Ward no.6”, “In the Ravine” лексически богаче британских “Dreary Story”, “Ward №6”, “In the Ravine”. Одно остается неизменным – все проанализированные тексты А. П. Чехова лексически вариативнее их англоязычных версий перевода. Можно предположить, что тексты оригиналов богаче переводов, поскольку переводчикам не всегда удается выразительно перевести оригинал на другой язык, при этом сохранив уникальный стиль и лексическую выразительность автора оригинала. Считается важным осуществление предварительного литературоведческого анализа творчества писателя для того, чтобы верно «интерпретировать имплицитные замыслы автора» [21, с. 292]. В случае переводов с русского на английский язык на уникальность лексической составляющей текстов влияет и сама структура языков, что вынуждает переводчика прибегать к разнообразным приемам при переводе лексических структур оригинала. Русский язык способен воспроизводить новую лексику путем изменений внутри самой лексемы, к примеру глаголы с одним корнем и разными префиксами (*приходить*, *входить*) воспринимаются программой как две разные леммы, тогда как в английском переводе глаголы (*come from*, *come in*) машина воспринимает как одну и ту же лемму *come*, при этом предлоги считаются отдельными леммами. Таким образом, синтетический русский язык отличается от аналитического английского по структуре, что влияет на частотность русских и английских лексем.

Взаимосвязь между уникальными лексическими единицами и всеми словоупотреблениями в оригиналах и переводах, как справедливо отмечают исследователи Л. Л. Гонсалвес и Л. Б. Гонсалвес [15, с. 561], можно представить в виде линейной регрессии, которая отражает сложную лингвистическую реальность в виде математической модели. В нашем случае линейная регрессия была построена на материале тридцати рассказов А. П. Чехова и версий их перевода, как показано в статье «Коэффициент лексического разнообразия в текстах оригинала и перевода» [19, с. 38]. В настоящем исследовании в соответствии со значениями показателей *N*, *n* и *k* графически представлена линейная регрессия по данным из оригиналов пьесы А. П. Чехова «Вишневый сад», его рассказов «Скучная история», «Палата №6», «В овраге», а также двадцати семи рассказов, отобранных методом сплошной выборки (рис. 1). Также представлены графики линейной регрессии по данным из англоязычных переводов тех же произведений А. П. Чехова, а именно пьесы «Вишневый сад», рассказов «Скучная история», «Палата №6», «В овраге»,

а также двадцати семи рассказов, осуществленных К. Гарнетт (рис. 2), Р. Нельсоном, Р. Пивером и Л. Волохонской (рис. 3).

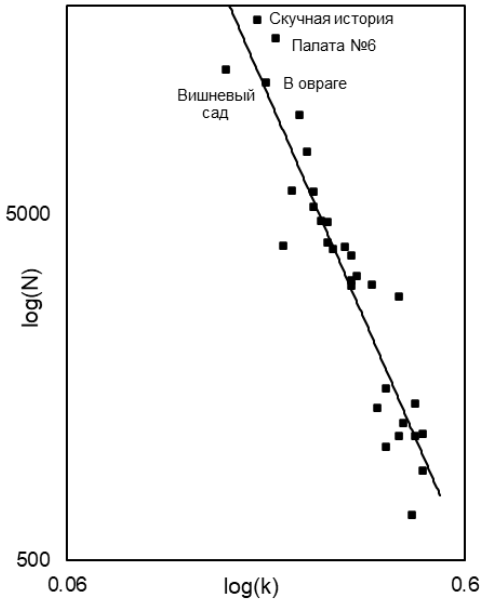

Рис. 1. Линейная регрессия по данным из пьесы А. П. Чехова «Вишневый сад», рассказов «Скучная история», «Палата №6», «В овраге» и других двадцати семи рассказов.

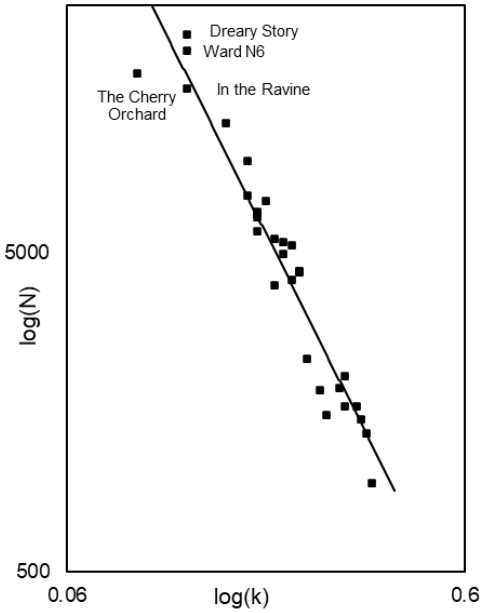

Рис. 2. Линейная регрессия по данным из переводов на английский язык пьесы А. П. Чехова «Вишневый сад» (“The CherryOrchard”), рассказов «Скучная история» (“Dreary Story”), «Палата №6» (“Ward №6”), «В овраге» (“In the Ravine”) и других двадцати семи рассказов, осуществленных К. Гарнетт.

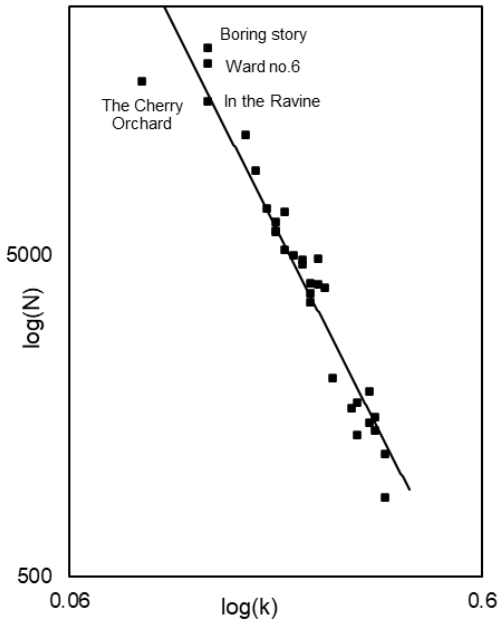

Рис. 3. Линейная регрессия по данным из переводов на английский язык пьесы А. П. Чехова «Вишневый сад» (“The Cherry Orchard”), рассказов «Скучная история» (“Boring Story”), «Палата №6» (“Ward no.6”), «В овраге» (“In the Ravine”) и других двадцати семи рассказов, осуществленных Р. Нельсоном, Р. Пивером и Л. Волохонской.

При сравнении полученных результатов становится понятно, что линейная регрессия для текстов А. П. Чехова больше смещена в сторону показателя $k = 0.6$, т.е. в сторону большего лексического разнообразия, в отличие от линейных регрессий британских и американских переводов. Обе линейные регрессии переводных текстов близки по значению, т.е. “The CherryOrchard” и тридцать рассказов находятся примерно, в одной и той же позиции. Более того, тексты А. П. Чехова («Вишневый сад», «Скучная история», «Палата №6», «В овраге») и аналогичные им британские и американские переводы расположены в верхней части линейной регрессии, при этом дальше от показателя $k = 0.6$, что наглядно демонстрирует взаимосвязь объема текста и его лексической вариативности, т.е. чем больше объем текста, чем меньше лексическое разнообразие. Что касается пьесы «Вишневый сад», то как в оригинале, так и в переводах пьеса смещена в сторону наименьшей лексической вариативности, т.е. ближе к показателю $k = 0.06$, что подтверждает выводы о том, что пьеса «Вишневый сад» менее лексически вариативна, чем некоторые из взятых для анализа тридцати рассказов А. П. Чехова.

Таким образом, в проведенном исследовании предлагается использовать методы автоматической обработки текстов для сопоставления лексической наполненности оригиналов и переводов произведений А. П. Чехова. Опираясь на методы лемматизации и частеречной разметки, можно прийти к выводу, что в пьесе А. П. Чехова «Вишневый сад», по сравнению с некоторыми его рассказами, глагольная лексика употребляется чаще, что обусловлено особой структурой драматургического текста, в котором глаголы

играют особенно важную роль. Кроме того, пьеса А. П. Чехова «Вишневый сад» лексически разнообразнее, чем оба ее англоязычных перевода. В то же время по сравнению с рассказами А. П. Чехова пьеса «Вишневый сад» менее лексически вариативна. То же самое можно сказать и об англоязычных переводах пьесы и рассказов. Наименьший по значению коэффициент лексического разнообразия пьесы А. П. Чехова «Вишневый сад» в сравнении с коэффициентом в его рассказах в определенной степени объясняется тем обстоятельством, что жанр пьесы предполагает наличие диалогов, монологов, реплик персонажей, которые изобилуют лексическими повторами, что, соответственно, снижает ее лексическое разнообразие. Напротив, широкое использование описательной лексики в рассказах закономерно способствует увеличению коэффициента лексического разнообразия. Построенная линейная регрессия позволяет наглядно продемонстрировать взаимосвязь между объемом текста и его уникальными лексическими единицами и сделать вывод о том, что тексты оригиналов А. П. Чехова лексически вариативнее версий их переводов. Результаты исследования могут быть использованы для дальнейшей разработки проблемы проведения сопоставительного анализа оригиналов и их переводов методами автоматической обработки текстов.

Литература

1. Мифтахова Р. Г., Морозкина Е. А. Нейронное представление семантического поля // Вестник Башкирского университета. 2021. Т. 26. №4. С. 1130–1135.
2. Мифтахова Р. Г., Морозкина Е. А. Машинный перевод. Нейроперевод // Вестник Башкирского университета. 2019. Т. 24. №2. С. 497–502.
3. MyStem: Программа MyStem производит морфологический анализ текста на русском языке. URL: https://yandex.ru/dev/mystem/
4. Laurence Anthony's Website: Software. URL: https://laurenceanthony.net/software.html
5. Чехов А. П. Вишневый сад. Драма на охоте. Дама с собачкой. Повести. Рассказы. М.: Эксмо, 2020. 1024 с.
6. Chekhov A. The Cherry Orchard and other plays translated by Garnett C. London: Heron Books, 1968. 430 p.
7. Chekhov A. The Cherry Orchard translated by Nelson R., Pevear R., Volokhonsky L. New York: Theatre Communications Group, 2015. 287 p.
8. Rusk J. Chekhov Stories in the Order of English Publication translated by Garnett C. URL: https://www.ibiblio.org/eldritch/ac/jr/garnett.htm
9. Chekhov A. Fifty-Two Stories translated by Pevear R., Volokhonsky L. New York: Vintage, 2020. 530 p.
10. Шафиков С. Г. Типология языков, метаязык лингвистической типологии и языковые универсалии // Доклады Башкирского университета. 2020. Т. 5. №6. С. 439–443.
11. Пешкова Н. П. Метод триангуаляции как инструмент полидисциплинарного подхода к исследованию речевой коммуникации // Доклады Башкирского университета. 2022. Т. 7. №5. С. 316–323.

12. Сафина З. М., Корнилова А. Д. Передача видовременных форм русских глаголов на английский язык // Доклады Башкирского университета. 2021. Т. 6. №2. С. 122–129.
13. Морозкина Е. А., Исхакова Э. В. Интерпретация «текстовых аномалий» в геральдической конструкции интертекстуальности // Доклады Башкирского университета. 2022. Т. 7. №6. С. 426–434.
14. Морозкина Е. А., Морозкин Ю. Н., Сафина З. М. Фрактальные свойства глаголов движения в оригинале и переводе художественного текста // Вестник башкирского университета. 2018. Т. 23. №3. С. 777–782.
15. Gonçalves L. L., Gonçalves L. B. Fractal power law in literary English // Physica A: Statis-tical Mechanics and its Applications. 2006. Vol. 360. Issue 2. Pp. 557–575.
16. Grieve J. Quantitative Authorship Attribution: An Evaluation of Techniques // Literary and Linguistic Computing. 2007. Vol. 22. No. 3. Pp. 251–270.
17. McCarthy Ph. M., Jarvis S. Voc-D: A theoretical and empirical evaluation // Language Testing. 2007. Vol. 24. No. 4. Pp. 459–488.
18. Бузаджи Д. М., Ланчиков В. К. Буквализм и языковое разнообразие. Об использовании одного метода корпусной лингвистики в переводоведении // Мосты. 2011. №4(32). С. 12–31.
19. Сафина З. М., Корнилова А. Д., Смакова А. Л. Коэффициент лексического разнообразия в текстах оригинала и перевода // Языки в диалоге культур: проблемы многоязычия в полиэтническом пространстве: мат-лы V Всерос. научно-практ. конф. с междунар. участием (г. Уфа, 5 мая 2022 г.) / отв. ред. А. С. Самигуллина. Уфа: РИЦ БашГУ. 2022. С. 37–40.
20. 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.
21. Морозкина Е. А., Тимирбаева О. О. Анализ перевода семантического парадокса (на мат-ле романа М. Ю. Лермонтова «Герой нашего времени» и англоязычных версий его перевода) // Доклады Башкирского университета. 2019.Т. 4. №3. С. 291–296.

Natural Language Processing of A. Chekhov's literary texts and of their English-language translation versions based on the methods of lemmatization and part-of-speech tagging

E. A. Morozkina, A. D. Kornilova*

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

**Email: anastasia.ufa@mail.ru*

The article offers to conduct a comparative analysis of the original literary texts and of their English-language translation versions using Natural Language Processing techniques in order to identify the

coefficient of lexical diversity. It is found out that A. Chekhov's play “The Cherry Orchard” and a number of his short stories are lexically more variable than their English-language translation versions. It turned out that in A. Chekhov's play “The Cherry Orchard” verbal lexical units are used more often than the units of other major parts of speech, while in A. Chekhov's short stories nouns prevail. An attempt is made to represent in the form of linear regression the dependence of the coefficient of lexical diversity on the volume of the text.

Keywords: lemmatization, part-of-speech tagging, natural language processing.