A Quantitative Study of Simplification Strategies in Adapted Texts for L2 Learners of Russian

Anna Dmitrieva
University of Helsinki, Finland
HSE University, Moscow, Russia
Pushkin State Russian Language Institute,
Moscow, Russia
annadmitrieva252@gmail.com

Antonina Laposhina
Pushkin State Russian Language Institute,
Moscow, Russia
antonina.laposhina@gmail.com

Maria Lebedeva
Pushkin State Russian Language Institute,
Moscow, Russia
m.u.lebedeva@gmail.com

Abstract

Nowadays there has been a growing interest in the topic of Russian text adaptation, both in theoretical aspects of intralingual translation into Simple and Plain Russian, and in practical tasks like automatic text simplification. Therefore, it is important to study the characteristics that make an adapted text more accessible. In this paper, we aim to investigate the strategies that human experts employ when simplifying texts, particularly when the texts are being adapted for learners of Russian as a foreign language. The main data source for this research is the RuAdapt parallel corpus, which consists of Russian literature texts adapted for the learners of RaaFL and the original versions of these texts. We study the changes that occur during the adaptation process on lexical, morphological, and syntax level, and compare them to the methods usually described in methodological recommendations for teaching RaaFL.

Keywords: simplification, simplified Russian, adaptation, adapted text, Russian as a foreign language, corpus of simplified texts

DOI: 10.28995/2075-7182-2021-20-191-203

Квантитативное исследование стратегий упрощения на материале адаптированных текстов для изучающих РКИ

Анна Дмитриева
Университет Хельсинки, Финляндия
НИУ ВШЭ, Москва, Россия
Гос. ИРЯ им. А. С. Пушкина,
Москва, Россия
annadmitrieva252@gmail.com

Антонина Лапошина
Гос. ИРЯ им. А. С. Пушкина,
Москва, Россия
antonina.laposhina@gmail.com

Мария Лебедева
Гос. ИРЯ им. А. С. Пушкина, Москва, Россия
m.u.lebedeva@gmail.com

Аннотация

В настоящее время возрастает интерес к теме упрощения текстов на русском языке: как к теоретическим аспектам внутриязычного перевода на простой и ясный русский, так и к практическим задачам, таким, как автоматическое упрощение. Таким образом, важным представляется изучить характеристики, делающие

адаптированные тексты более доступными. Целью данной работы является изучение стратегий, применяемых экспертами при упрощении текста, в особенности при упрощении для изучающих русский язык как иностранный. Основным источником данных для нашего исследования является параллельный корпус RuAdapt, включающий в себя тексты русской литературы, адаптированные для изучающих РКИ, и их оригинальные версии. Мы изучаем изменения, которые можно наблюдать в процессе адаптации на лексическом, морфологическом и синтаксическом уровне, и сравниваем их с методами, которые часто описывают в методических рекомендациях для преподавания РКИ.

Ключевые слова: упрощение, упрощенный русский язык, адаптация, адаптированный текст, русский язык как иностранный, корпус упрощенных текстов

1 Введение

Задача автоматического упрощения текста является одной из актуальных и нетривиальных задач в области обработки естественного языка. Концепция упрощенного русского языка находится в настоящий момент на этапе становления; при этом исторически наиболее разработанной областью является упрощение языка в учебных целях, или учебная адаптация текста.

Адаптированный текст – результат специальной обработки аутентичного текста, проведённой с опорой на определенные дидактические принципы. В практике обучения иностранному языку такие принципы определяются в соответствии с требованиями к владению языком на определённом уровне и в соответствии с тем, насколько текст «полезен» в учебном плане. Наиболее продуктивными стратегиями адаптации признаются упрощение лексических и синтаксических структур [14, с.94], а также опущение фраз или предложений. Однако в соответствии с принципом дидактической целесообразности для адаптации могут использоваться стратегии, противоположные упрощению: например, текст может специально насыщаться изучаемыми лексическими или грамматическими единицами, фрагменты текста могут объясняться и таким образом становиться пространнее и структурно сложнее [35, с. 269]. Адаптация текстов на русском языке для иностранных учащихся также опирается на базовые стратегии замены всех компонентов, затрудняющих восприятие текста, исключения несущественных компонентов и добавления комментариев [9].

Таким образом, адаптация текста представляет собой сложный процесс, базирующийся, с одной стороны, на типичных стратегиях упрощения, с другой стороны, на специфических стратегиях и требованиях к тексту.

На материале русского языка проблема адаптации текста с позиций компьютерной лингвистики исследовалась в работах [7] [15]. Так, в исследовании [34] изучались стратегии адаптации, которыми пользовались преподаватели русского языка как иностранного (РКИ) при упрощении новостных текстов. Было показано, что адаптированные тексты отличаются от оригиналов рядом лексических, морфологических и семантических особенностей.

Вклад в понимание того, какие стратегии используются при адаптации текстов для разных целей, способно внести исследование специальных представительных датасетов, состоящих из выравненных пар оригинальных и адаптированных текстов. Существует ряд корпусов подобного типа на материале английского [17] [26], французского [14], испанского [38] языков. На материале русского языка такие корпуса в настоящий момент только разрабатываются. Для решения задач данного проекта был создан параллельный корпус RuAdapt, в состав которого вошли адаптированные художественные тексты, предназначенные для изучающих РКИ, и их оригиналы. Цель данной работы состоит в сравнительном исследовании пар оригинальных и адаптированных текстов количественными методами и формализации описываемых в методической литературе стратегий учебной адаптации текста. Такое исследование способно дополнить и уточнить представления о том, какие формальные критерии определяют упрощенный русский язык, и внести вклад в решение задачи автоматической симплификации текстов на русском языке.

2 Материалы и методы

2.1 Данные

Корпус RuAdapt[1] содержит полные тексты оригинальных текстов и их адаптированных версий. Объём адаптированных текстов в настоящий момент составляет 268 тыс. словоупотреблений. Каждый адаптированный текст (за исключением небольшого числа малоизвестных современных рассказов) имеет оригинальную пару; объем оригинальных текстов составляет 885 тыс. словоупотреблений. Тексты в датасете выровнены по параграфам. На данный момент большая часть датасета находится в открытом доступе, за исключением тех произведений, оригинальные версии которых еще не перешли в общественное достояние либо не были опубликованы автором в свободном доступе. Тексты были выровнены автоматически с использованием нескольких элайнеров: Bleualign[2] [31] и CATS[3] [36].

Большая часть корпусами представлена произведениями русской классической литературы от рассказа до романа (А.П. Чехов, Л.Н. Толстой и пр.), однако есть и некоторые произведения современных авторов (Б. Акунин, В. Токарева). Основной объем корпуса составили тексты, предназначенные для уровня B1 (см. Рис. 1).

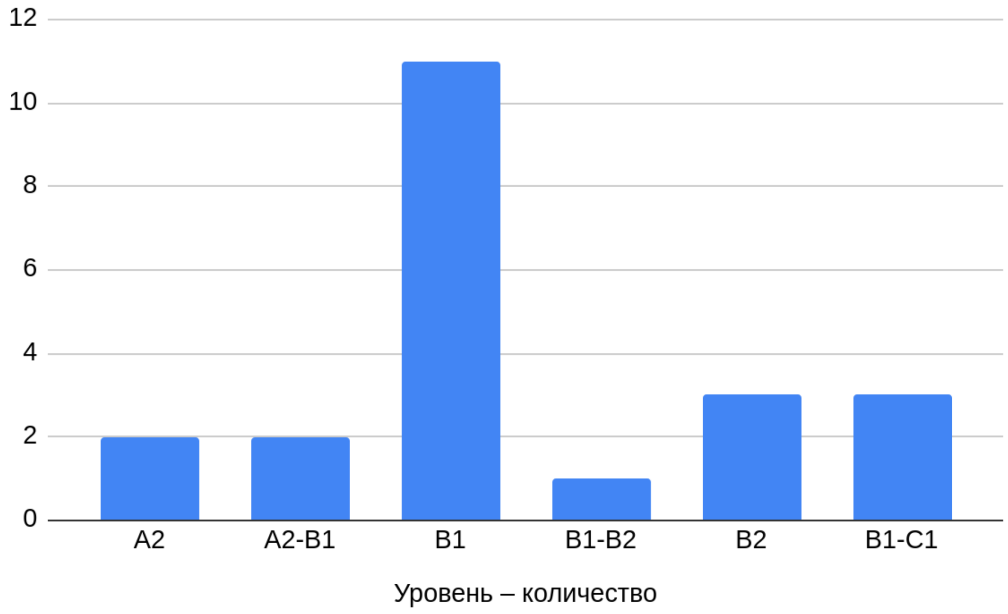

Рис.1. Распределение текстов разных уровней владения РКИ в корпусе RuAdapt

Указание нескольких уровней сразу (B1-B2) используется, если книга предназначена для переходного этапа между этими уровнями, либо когда диапазон уровней дан для всего сборника рассказов, и определить уровень каждого рассказа в отдельности невозможно.

Данное исследование проводилось на материале полных текстов произведений.

2.2 Методы

На этапе предобработки тексты были распознаны из формата PDF в обычный текстовый формат при помощи Apache Tika[4]. Затем и оригинальные, и адаптированные тексты были очищены от шума (ударений, ненужных пробелов и т.д.) автоматически при помощи Python 3.

Для оценки изменений, происходящих с оригинальными текстами в процессе адаптации, были рассчитаны показатели, традиционно применяющиеся для задач автоматического определения

[1] https://github.com/Digital-Pushkin-Lab/RuAdapt

[2] https://github.com/rsennrich/Bleualign

[3] https://github.com/neosyon/SimpTextAlign

[4] https://tika.apache.org/

сложности русского текста [16] [21] [29] [33], которые условно можно разделить на лексические (процент покрытия текста лексическими минимумами, частотными списками и др.); морфологические (количество различных частей речи и грамматических форм); синтаксические (глубина глагольных и именных групп, связи между глаголами в предложениях); признаки, основанные на базовых подсчетах (средняя длина слов и предложений, а также различные метрики удобочитаемости).

Для токенизации и лемматизации был использован Mystem 3 [30] – токенами признавались только слова, выделяемые Mystem, таким образом, пунктуация и цифры отдельными токенами не считались. Для сегментации предложений был использован модуль ru_punkt[5], в настоящий момент встроенный в библиотеку NLTK в Python. Для исследования синтаксических характеристик текстов использовались возможности библиотеки deeppavlov[6], синтаксический парсер в которой позволяет производить полный разбор предложений по схеме Универсальных зависимостей (Universal Dependencies) и получать выдачу в формате CONLL.

3 Результаты

3.1 Исследование характеристик текстов

Чтобы получить общее представление о характеристиках оригинальных и адаптированных текстов, было решено рассмотреть морфологические, лексические и синтаксические характеристики оригинальных и адаптированных текстов.

Признак	**Оригинальные тексты**	**Адаптированные тексты**
Среднее кол-во слов в тексте	6190	1877
Среднее кол-во предложений	488	203
Средняя длина слова в слогах*	2.04	1.97
Средняя длина слова*	5.08	4.89
Средняя длина предложения*	12.85	9.66
Среднее количество пунктуации на предложение*	2.4	1.7
Индекс SMOG	10.52	9.0
Индекс Dale-Chale	9.85	8.44
Flesch-Kincaid Grade Level (FKGL)	3.03	1.59
Flesch Reading Ease	65.5	71.53
Индекс Coleman-Liau	4.87	3.38
Automated Readability Index (ARI)	4.89	3.34

Таблица 1. Общие сходства и различия оригинальных и адаптированных текстов

В Таблице 1 представлены сходства и различия оригинальных и адаптированных текстов по некоторым общим параметрам. Признаки, отмеченные знаком *, являются взвешенными

[5] https://github.com/Mottl/ru_punkt

[6] https://github.com/deepmipt/DeepPavlov

средними: например, средняя длина слова во всех текстах считается как среднее средних длин слов в каждом тексте, взвешенное на количество слов в данном тексте.

Для изучения сложности словарного состава текстов мы также обратили внимание на различные метрики удобочитаемости, такие как индекс SMOG[23], индекс ARI [32], Flesch Reading Ease [37], Flesch-Kincaid Grade Level [20], Coleman-Liau [11], Индекс Dale-Chale [13]. Константы для формулы Флеша адаптированы для русского языка И.В.Оборневой [25], для остальных метрик — И. Бегтиным[7]. В таблице 1 можно найти значения некоторых из использованных нами метрик удобочитаемости. Значение метрик удобочитаемости рассчитывалось отдельно для каждого текста.

Из таблицы 1 видно, что при адаптации тексты чаще всего сильно сокращаются. При этом средние длины слов в символах и слогах меняются не очень сильно, но предложения становятся короче и проще, если судить по количеству пунктуации. Также можно видеть, что удобочитаемость адаптированных текстов в среднем выше, однако это различие не всегда велико. Поскольку в исследовании используется несколько индексов удобочитаемости, все индексы были попарно сравнены между собой с использованием t-критерия Стьюдента для двух выборок. Почти все пары, кроме двух (Индекс Dale-Chale и Индекс SMOG, Индекс Coleman-Liau и ARI), имеют существенные статистически значимые различия.

3.2 Лексический уровень адаптации

Упрощение лексики текста является одним из самых очевидных и ключевых направлений адаптации текста в учебных целях. Так, многочисленные исследования говорят о самой тесной связи знакомости лексики текста и успешности его понимания [24] [28].

3.2.1 Лексические минимумы

Одним из наиболее разработанных показателей доступности лексики текстов, предназначенных для изучающих РКИ, является покрытие текста лексическими минимумами – специальными списками слов, которые студент должен знать в зависимости от уровня владения языком по шкале CEFR, от A1 до C1. При подсчете вхождений лексики из текстов в минимумы для различных уровней освоения РКИ имена, фамилии, отчества и географические названия считались знакомыми читателю словами. Для расчетов использовалась классическая линейка лексических минимумов системы ТРКИ [1-5]. Лексика, которой нет в минимуме для уровня C1 (11% для оригинальных текстов и 7% для адаптированных), считается не покрытой лексическими минимумами.

Список	Объем списка	Оригинальные тексты, %	Адаптированные тексты, %
A1	900	58	63
A2	1300	66	72
B1	2300	74	81
B2	5500	82	87
C1	11000	89	93

Таблица 2. Покрытие текстов корпуса лексическими минимумами разных уровней

Как можно видеть из таблицы 2, процент слов из лексических минимумов, знакомых потенциальному читателю, закономерно оказывается выше в адаптированных текстах. Однако в

[7] https://github.com/infoculture/plainrussian

некоторых случаях эта разница не слишком существенна, и даже на уровне C1 не все слова в адаптированных текстах в среднем знакомы аудитории. Это может объясняться доменной спецификой корпуса: многие произведения русской классической литературы достаточно сложны для чтения и в большинстве случаев не могут быть полностью адаптированы.

3.2.2 Частотные списки слов

Частотность слова также является значимым критерием доступности лексики и традиционно учитывается при адаптации текстов. Замена редких слов на более частотные синонимы нередко применяется в системах лексического упрощения для снижения сложности текста [10][18].

Список	**Оригинальные тексты, %**	**Адаптированные тексты, %**
Частотный список 1 000	49	54
Частотный список 3 000	65	70
Частотный список 5 000	72	77
Частотный список 10 000	81	84

Таблица 3. Покрытие текстов корпуса частотными списками

Таблица 3 содержит данные о покрытии текстов корпуса списками самых частотных слов русского языка по Новому частотному словарю русской лексики (далее – Частотный словарь) [22]. Процент частотной лексики стабильно выше в адаптированных версиях, что соответствует основным стратегиям лексической адаптации текстов.

3.2.3 Стратегии лексической адаптации

Для того чтобы проиллюстрировать изменения лексического состава текстов в процессе адаптации на реальных примерах, был проведен сравнительный анализ частотных списков оригинальных текстов и их адаптированных версий. Для поиска лексики, максимально отличающейся по частотности в адаптированных версиях текста, был использован рейтинг ключевых слов (keyness score) [19]. На основании этого анализа были отмечены следующие стратегии адаптации.

Поскольку корпус содержит большое количество текстов русской классической литературы, большая доля изменений в лексике связан с работой с устаревшей лексикой (пункты 1-3). Остальные стратегии более универсальны и так или иначе встречаются во всех текстах коллекции. В Таблице 4 приведены значения встречаемости отдельных слов в оригинальных и адаптированных версиях, а также частотность слов по Частотному словарю.

1. Замена устаревшего слова на современный аналог (*нынче – сегодня; подле – у,*) или вариант написания (*чрез – через, кофий – кофе*).

2. Замена историзма на синоним (*лакей, человек* в значении прислуги – *слуга*; *гусар – офицер*)

3. Удаление слова без передачи смысла другими словами (*кучер, земский*)

Так, пример 1 демонстрирует все 3 перечисленные стратегии адаптации лексики произведения классической литературы.

(1) a. *Дуня села в кибитку подле гусара, слуга вскочил на облучок, ямщик свистнул, и лошади поскакали.*

b. *Дуня села рядом с офицером, и они поехали.*'

4. Удаление слова или сочетания и передача смысла другими словами. Эту стратегию достаточно трудно обнаружить с помощью сравнения контекстов, поскольку смысл может быть передан самыми различными средствами.

(2) a. — *Она здорова, —* ***хмурясь промычал*** *Алексей Александрович.*

b. — *Она здорова, —* ***недовольно ответил*** *Алексей Александрович.*

5. Замена слова на более частотный синоним или гипероним (повесить – убить, промычать – сказать).

6. Замена слова с суффиксами субъективной оценки (дверца – дверь, мальчишка– мальчик).

7. Полная переработка предложения (пример 3). Данный вид изменения текста также сложно найти с помощью сравнения частотных списков, поэтому трудно судить о количестве подобных примеров.

(3) a. *Гав, говорю, идиотка!*

b. *Я, конечно, обиделся.*

Стратегия адаптации	**Лемма**	**Частотность по корпусу оригинальных текстов (ipm)**	**Частотность по корпусу адаптированных текстов (ipm)**	**Частотность по Частотному словарю (ipm)**
1	дурной	150	5 (!)[8]	31
	плохой	66	196	222
1	увидать	314	126	7
	увидеть	346	1058	452
2	лакей	187	23	5
	слуга	156	69	18
3	земский	40	0	5
3	пухлый	51	0	10
3, 4	кучер	108	5 (!)	4
4	хмуриться	40	0	6
5	почтительный	50	9 (!)	4
	уважительный	8	23	7
6	мальчишка	41	22	56
	мальчик	170	149	188

Таблица 4. Стратегии учебной адаптации в художественных текстах

[8] Знаком (!) обозначены слова, встретившиеся в корпусе менее 3 раз

3.3 Морфологические характеристики

В таблице ниже приведены средние относительные (к объему текста в словах) частоты некоторых частей речи. Относительные частоты подчинительных и сочинительных союзов были подсчитаны на основе морфологических разборов deeppavlov, частоты остальных частей речи – на основе морфологического разбора Mystem.

3.3.1 Относительные частоты частей речи

Часть речи	Оригинальные тексты	Адаптированные тексты
Существительное (S)	0.26	0.25
Глагол (V)	0.17	0.17
Сочинительный союз (CCONJ)	0.05	0.05
Подчинительный союз (SCONJ)	0.02	0.02
Прилагательное (A)	0.07	0.06
Наречие (ADV)	0.06	0.06
Числительное (NUM)	0.008	0.009

Таблица 5. Относительные частоты частей речи

Видно, что относительные частоты частей речи не сильно меняются от оригинальных версий к адаптированным. Тем не менее, изменяются некоторые морфологические характеристики данных частей речи. Так, изучение частот различных глагольных форм на основании разборов deeppavlov показывает, что относительная частота финитных глаголов в адаптированных текстах повышается, а деепричастий (Conv) и причастий (Part) – снижается. Количество инфинитивов при этом остается неизменным.

Глагольная форма	Оригинальные тексты	Адаптированные тексты
Финитные глаголы	0.74	0.77
Инфинитивы	0.14	0.15
Причастия	0.07	0.05
Деепричастия	0.05	0.03

Таблица 6. Средняя относительная частота глагольных форм по отношению ко всем глаголам

Кроме этого, в адаптированных текстах повышается относительное количество глаголов в изъявительном наклонении (с 0.72 до 0.75) и уменьшается относительное количество прилагательных в полной форме (с 0.99 до 0.93). Снижается также количество имен в творительном падеже: от 0.07 в оригинальных текстах до 0.06 в адаптированных.

3.3.2 Синтаксические характеристики

Заметно, что максимальные и средние глубины глагольных и именных групп существенно снижаются в адаптированных текстах, что может косвенно указывать на наличии более простых предложений в адаптации. Именными группами считались группы, где вершиной является существительное, личное местоимение или имя собственное.

Признак	Оригинальные тексты	Адаптированные тексты
Максимальная глубина глагольной группы	46.19	28.8
Средняя глубина глагольной группы	7.71	6.33
Максимальная глубина именной группы	27.90	18.72
Средняя глубина именной группы	3.52	3.12

Таблица 7. Глубины групп

Говоря о связях внутри глагольных групп, можно отметить незначительное сокращение сочинительных связей (conj), а также субъектов клауз, в т.ч. пассивных (csubj, csubj:pass). Сокращается также количество наречий-модификаторов, в том числе модификаторов клауз (advcl, advmod). При этом повышается количество различных дополнений клауз (xcomp, ccomp), а также паратаксиса. Таким образом, можно сделать вывод о том, что синтаксические структуры в адаптированных текстах становятся более простыми, хотя и не слишком упрощаются, судя по количеству дополнений клауз. Паратаксис также может свидетельствовать о сохранении предложений с прямой речью.

Тип связи в UD	Оригинальные тексты	Адаптированные тексты
Open clausal complement (xcomp)	0.17	0.2
Adverbial clause modifier (advcl)	0.16	0.14
Conjunct (conj)	0.47	0.44
Parataxis	0.12	0.13
Clausal complement (ccomp)	0.06	0.08
Clausal subject (csubj)	0.013	0.012
Clausal subject – passive (csubj:pass)	0.0016	0.0019
Adverbial modifier (advmod)	0.0006	0.0003

Таблица 8. Относительные частоты связей внутри глагольных групп

Можно проследить изменения в синтаксических структурах на следующем примере:

(4) a. *Анвар даже пробовал выговорить доблестному злодею помилование, но озлобившиеся министры были непреклонны, и наутро убийцу повесили на дереве. Дамы из гарема, так горячо любившие своего Черкеса, пришли посмотреть на его казнь, горько плакали и посылали ему воздушные поцелуи.*

b. *Когда эфенди узнал о том, что случилось, он просил министров не быть слишком жестокими к его другу. Но министры его даже слушать не стали. Утром Гасана убили. Женщины во дворце плакали.*

Видно, что была упразднена сочинительная связь в первом предложении в пользу нескольких простых предложений. Также исключаются причастия и адвербиальные модификаторы (так горячо любившие), при этом замена не осуществляется. С разделением сложных предложений снижается также глубина глагольных и именных групп.

3.4 Статистическое тестирование и моделирование зависимости между классом текста и его характеристиками

Для изучений возможных зависимостей между классом текста (оригинал или адаптированный) был применен коэффициент ранговой корреляции Кендалла. В результате исследования было выяснено, что наибольшую отрицательную корреляцию с классом текста имеют такие метрики, как процент слов, входящих в лексические минимумы для ТРКИ, некоторые лексические списки, а также значение формулы Оборневой. Наибольшую же положительную корреляцию имеют такие признаки, как некоторые формулы удобочитаемости, относительная частота причастий и деепричастий, максимальная и средняя глубина глагольной группы, количество устаревших слов, слов в творительном падеже и некоторые типы связей внутри глагольной группы (advmod, obj).

Для дальнейшего исследования зависимости между признаками и классом текста была построена логистическая регрессия. Для построения данной модели использовалась библиотека sklearn [27]. Перед подачей в модель тексты были перемешаны, размер тестового множества составил 0.2 от всей выборки. Также перед подачей в модель значения признаков были масштабированы от 0 до 1. Регрессия использовалась с параметрами по умолчанию, для оптимизации был выбран метод покоординатного спуска (liblinear solver). F1-мера классификации составила 71 для оригинальных текстов и 70 для класса адаптированных текстов. Это позволяет говорить о способности признаков объяснять класс, к которому принадлежит текст.

При изучении признаков, имеющих наибольшую значимость в модели, было обнаружено, что в решении определения текста к классу адаптированных текстов большую роль снова играют такие признаки, как количество слов из лексических минимумов и некоторых списков лексики. Кроме того, влияние оказывает количество фамилий в тексте, количество существительных. На принадлежность к классу оригинальных текстов указывают такие признаки, как процент длинных слов (т.е. слов длиннее 3 слогов), некоторые формулы удобочитаемости (формула Флеша в адаптации Оборневой, Индекс Dale-Chale и SMOG), количество сочинительных союзов, а также максимальная глубина именных и глагольных групп.

4 Выводы

Изучив результаты анализа характеристик оригинальных и адаптированных текстов, можно прийти к нескольким выводам. Во-первых, общее сокращение объемов текстов, а также сокращение длин предложений и их сложности (количества пунктуации, глубины глагольных и именных групп) свидетельствует о том, что одной из основных стратегий упрощения является саммаризация. При этом она происходит не только на уровне удаления целых отрывков произведения, но и на уровне предложений. Кроме этого, количественное исследование позволило подтвердить применение ряда описанных в литературе стратегий адаптации на морфологическом и синтаксическом уровне, в частности, замены деепричастий и причастий, замены сложных предложений несколькими простыми и пр.

Наиболее заметным на проанализированном материале оказывается редукция лексической сложности в процессе адаптации, о которой можно судить исходя из снижения процента редких слов и слов, выходящих за пределы лексических минимумов. При этом можно заметить, что некоторые стратегии упрощения, обнаруженные в других исследованиях, например, замена аббревиатур, очень мало представлены в исследуемом корпусе, вероятно, из-за специфики домена.

Другим интересным наблюдением является то, что при адаптации на лексическом уровне в упрощенных текстах заметно повышается количество слов из частотных списков. Это свидетельствует о том, что подобные списки (так же, как и лексические минимумы) можно использовать для автоматического лексического упрощения текстов. Использование частотных списков в этой задаче является распространенной практикой [18], которая, однако, на материале русского языка еще не применялось.

Для дальнейшего изучения стратегий упрощения необходимо будет сопоставить предложения из оригинальных текстов с предложениями из адаптированных, чтобы изучить, как именно происходят замены и/или удаления отдельных слов и фрагментов текста. Кроме этого, для создания дополнительных материалов для изучения стратегий упрощения, применяемых экспертами, будут привлечены эксперты-преподаватели РКИ.

Важным результатом данного исследования стал параллельный датасет RuAdapt, который может быть использован не только для изучения стратегий упрощения русского языка, подобного проведенному в настоящем исследовании, но и для создания либо дообучения систем автоматического упрощения текста. Поскольку в настоящее время задача автоматического упрощения часто рассматривается как монолингвальный машинный перевод [38], параллельные датасеты, где “обычным” сегментам соответствуют их упрощенные варианты, оказываются необходимы для обучения нейронных сетей для перевода. Кроме того, данные о стратегиях упрощения, полученные на основе подобных параллельных датасетов, можно применять для улучшения методов оценки результатов автоматического упрощения.

Понадобятся дополнительные исследования, чтобы сказать, насколько RuAdapt может улучшить качество нейронных моделей для упрощения, будучи использован в паре с другими датасетами, не относящимися к домену художественной литературы. Большую часть датасета составляют классические литературные произведения, актуальные для русского читателя по сей день. К тому же тексты, применяемые в учебных целях и упрощенные для читателей, осваивающих язык, нередко становятся частью параллельных датасетов для упрощения [6][8]. Это позволяет предположить возможность успешного использования RuAdapt с датасетами, содержащими общую лексику (например, такими, как датасет соревнования RuSimpleSentEval).

Благодарности

Работа выполнена с использованием средств государственного бюджета по госзаданию на 2020–2024 годы (проект FZNM-2020-0005).

References

[1] Andryshina N.P., Kozlova T.V. Lexical minimum of Russian as a foreign language. Level A1. Common language (4th ed.). – St. Petersburg, Zlatoust. 80 p. – 2012.

[2] Andryshina N.P., Kozlova T.V. Lexical minimum of Russian as a foreign language. Level A2.Common language (5th ed.). – St. Petersburg, Zlatoust. 116 p. – 2015.

[3] Andryshina N.P. (ed.) Lexical minimum of Russian as a foreign language. Level B1. Common language (9th ed.). – St. Petersburg, Zlatoust. 200 p. – 2017 (a).

[4] Andryshina N.P. (ed.), Lexical minimum of Russian as a foreign language. Level B2. Common language (7th ed.) – St. Petersburg, Zlatoust. 164 p. – 2017 (b).

[5] Andryshina N.P. (ed.). Lexical minimum of Russian as a foreign language. Level C1. Common language. – St. Petersburg, Zlatoust. 201 p. – 2018.

[6] Arfé B., Oakhill J., Pianta E. The text simplification in TERENCE //Methodologies and Intelligent Systems for Technology Enhanced Learning. – Springer, Cham, 2014. – P. 165-172.

[7] Baranova Yu. N., Elipasheva T. S. Creating an informational resource for Russian learner text analysis. [Sozdanie vspomogatel'nogo informacionnogo resursa dlya analiza uchebnyh tekstov na russkom yazyke.] // Chelovek v informacionnom prostranstve, Yaroslavl'. – 2014. – P. 232-246.
[8] Brouwers L. et al. Syntactic sentence simplification for French //Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR). – 2014. – P. 47-56.
[9] Brygina A.V. Linguistic principles of fiction text adaptation [Lingvisticheskie principy adaptirovaniya hudozhestvennogo teksta] // Ph.D. dissertation synopsis, RUDN university, Russia. – 2005. Available at: https://search.rsl.ru/ru/record/01003298898
[10] Chen X., Meurers D. Characterizing text difficulty with word frequencies //Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. – 2016. – P. 84-94.
[11] Coleman M., Liau T. L. A computer readability formula designed for machine scoring //Journal of Applied Psychology. – 1975. – Vol. 60. – №. 2. – P. 283.
[12] Crossley S. A., Yang H. S., McNamara D. S. What's so Simple about Simplified Texts? A Computational and Psycholinguistic Investigation of Text Comprehension and Text Processing //Reading in a Foreign Language. – 2014. – Vol. 26. – №. 1. – P. 92-113.
[13] Dale E., Chall J. S. A formula for predicting readability: Instructions //Educational research bulletin. – 1948. – P. 37-54.
[14] Gala N., Tack A., Javourey-Drevet L., Francois T., Ziegler J.C. Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers //Language Resources and Evaluation for Language Technologies (LREC). – 2020. – P. 1353-1361.
[15] Karpov N., Sibirtseva V., Bogdanov D., Dmitrieva A., Elian E., Kleshnin E., Markiva E., Teplukhina T., Violentova L. Development of modern electronic textbook of Russian as a foreign language: content and technology //Higher School of Economics Research Paper No. WP BRP. – 2012. – T. 6. Available at: https://ideas.repec.org/p/hig/wpaper/06hum2012.html.
[16] Karpov N., Baranova J., Vitugin F. Single-sentence readability prediction in Russian //International Conference on Analysis of Images, Social Networks and Texts. – Springer, Cham, 2014. – P. 91-100.
[17] Kauchak D. Improving text simplification language modeling using unsimplified text data //Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers). – 2013. – P. 1537-1546.
[18] Keskisärkkä R. Automatic text simplification via synonym replacement. – 2012. Ph.D. dissertation. Linköping University, Sweden, available at: https://www.diva-portal.org/smash/get/diva2:560901/FULLTEXT01.pdf
[19] Kilgarriff A. Simple maths for keywords //Proceedings of the Corpus Linguistics Conference. Liverpool, UK. – 2009. Available at: http://ucrel.lancs.ac.uk/publications/cl2009/171_FullPaper.doc
[20] Kincaid J. P. et al. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. – Naval Technical Training Command Millington TN Research Branch, 1975.
[21] Laposhina A. N., Veselovskaya T. S., Lebedeva M. U., Kupreshchenko O. F. Automated Text Readability Assessment For Russian Second Language Learners // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018". – Issue 17 (24). – 2018. Available at: http://www.dialog-21.ru/media/4312/laposhina%D0%B0n.pdf
[22] Lyashevskaya O.N., Sharov S.A. Modern Russian frequency dictionary (based on the data from the Russian National Corpus) [Chastotnyj slovar' sovremennogo russkogo yazyka (na materialah Nacionalnogo korpusa russkogo yazyka)] // Azbukovnik, Moscow. – 2009.
[23] Mc Laughlin G. H. SMOG grading-a new readability formula //Journal of reading. – 1969. – Vol. 12. – №. 8. – P. 639-646.
[24] Nation I. How large a vocabulary is needed for reading and listening? //Canadian modern language review. – 2006. – Vol. 63. – №. 1. – P. 59-82.
[25] Oborneva I. V. Automatic evaluation of text perception quality. [Avtomatizaciya ocenki kachestva vospriyatiya teksta] // Vestnik Moskovskogo gorodskogo pedagogicheskogo universiteta. Seriya: Informatika i informatizaciya obrazovaniya, (5). – 2005. – P. 86-91.
[26] Pavlick E., Callison-Burch C. Simple PPDB: A paraphrase database for simplification //Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). – 2016. – P. 143-148.

[27] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weis, R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: Machine learning in Python // The Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825-2830.

[28] Qian D. D. Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective //Language learning. – 2002. – Vol. 52. – №. 3. – P. 513-536.

[29] Reynolds R. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories //Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. – 2016. – P. 289-300.

[30] Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine //MLMTA. – 2003. – Vol. 2003. – P. 273.

[31] Sennrich R., Volk M. MT-based sentence alignment for OCR-generated parallel texts. – 2010.

[32] Smith E. A., Senter R. J. Automated readability index //AMRL-TR. Aerospace Medical Research Laboratories (US). – 1967. – P. 1-14.

[33] Sharoff S. K. S., Hartley A. Seeking needles in the web haystack: Finding texts suitable for language learners //8th Teaching and Language Corpora Conference. TaLC-8. – 2008.

[34] Sibirtseva V. G., Karpov N.V. Automatic adaptation of the texts for electronic textbooks. Problems and perspectives (on an example of Russian). [Avtomaticheskaya adaptaciya tekstov dlya elektronnyh uchebnikov. Problemy i perspektivy (na primere russkogo yazyka)] // Nová rusistika. – Vol. VII, číslo 1 – 2014. – P. 19-33.

[35] Siddharthan A. A survey of research on text simplification //ITL-International Journal of Applied Linguistics. – 2014. – Vol. 165. – №. 2. – P. 259-298.

[36] Štajner S. et al. Sentence alignment methods for improving text simplification systems //Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). – 2017. – P. 97-102.

[37] Flesch R. A new readability yardstick //Journal of applied psychology. – 1948. – T. 32. – №. 3. – P. 221.

[38] Xu W., Callison-Burch C., Napoles C. Problems in current text simplification research: New data can help //Transactions of the Association for Computational Linguistics. – 2015. – Vol. 3. – P. 283-297.