Квантитативное исследование стратегий упрощения на материале адаптированных текстов для изучающих РКИ

Анна Дмитриева[1,2,3], Антонина Лапошина[1], Мария Лебедева[1]

[1]ГосИРЯ им. А.С. Пушкина, [2]НИУ ВШЭ, [3]Университет Хельсинки

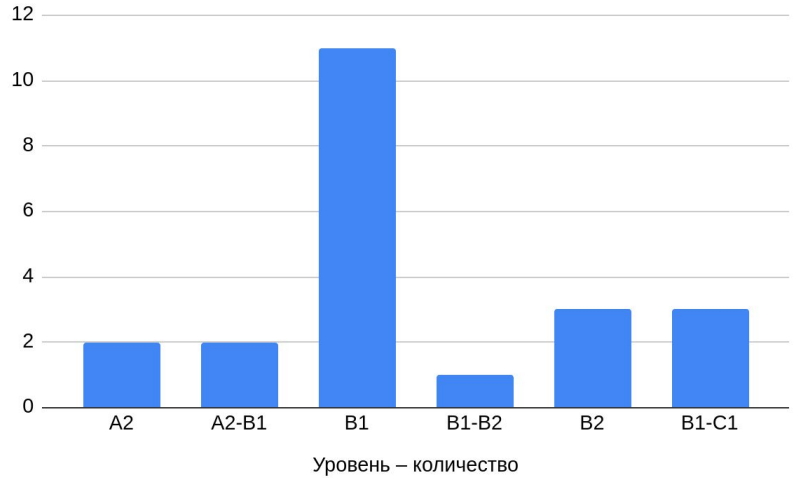
Цели и задачи исследования

- Изучить стратегии адаптации текстов на основе параллельного корпуса упрощенного русского языка.
- Определить формальные критерии, определяющие упрощенный русский язык. Как они соотносятся с тем, что описано в методической литературе?

RuAdapt: состав корпуса

Адаптированные художественные тексты, предоставленные издательством Златоуст, и их оригиналы, собранные из источников в открытом доступе.

Большинство текстов адаптированы для уровня B1.



Предобработка текстов

- Извлечение адаптированных текстов из PDF при помощи Apache Tika;
- Удаление шума, лишних пробелов и абзацев из адаптированных текстов и оригиналов;
- **Перед сбором статистики:** сегментация предложений (NLTK sent_tokenize), лемматизация (MyStem) в некоторых случаях, синтаксический анализ (deeppavlov).

Немного статистики

Метрика	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

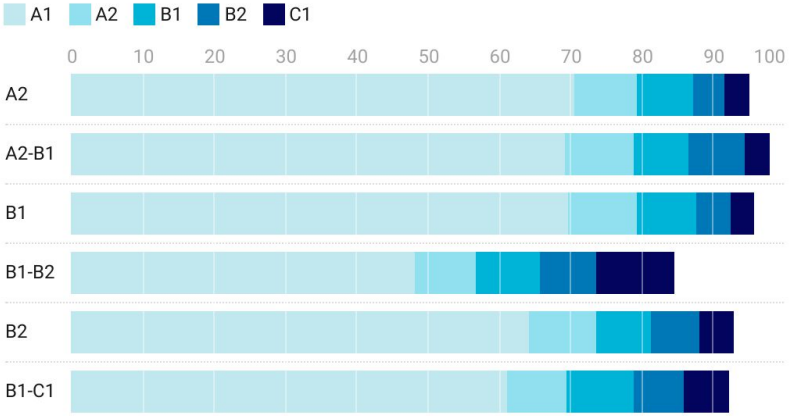
Удобочитаемость

Метрика	Оригинальные тексты	Адаптированные тексты
Индекс SMOG	10.52	9.0
Индекс Dale-Chale	9.85	8.44
Flesch-Kincaid Grade Level (FKGL)	3.03	1.59
Формула Флеша (в адаптации Оборенвой)	65.5	71.53
Индекс Coleman-Liau	4.87	3.38
Automated Readability Index (ARI)	4.89	3.34

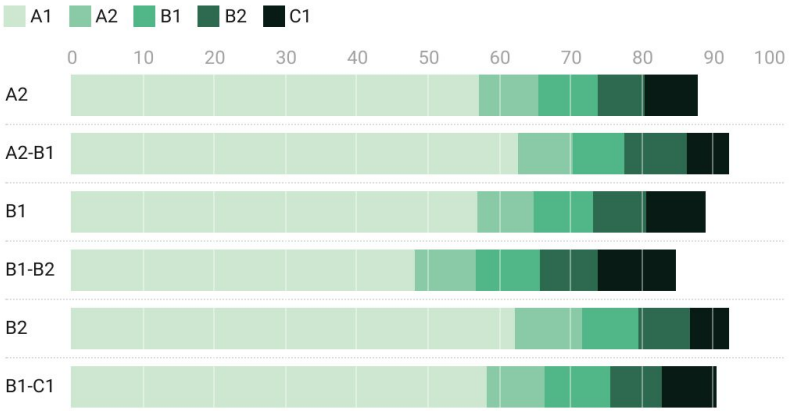
Вхождение слов в лексические минимумы

- Проверка, какой процент слов в текстах входит в лексические минимумы ТРКИ, т.е. является знакомым читателю на определенном уровне;
- Слова лемматизируются, собственные имена (фамилии, отчества) и географические названия считаются знакомыми читателям

Адаптированные тексты



Оригинальные тексты



Лексическая адаптация

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*).
4. Удаление слова или сочетания и передача смысла другими словами.

   a. — Она здорова, — **хмурясь промычал** Алексей Александрович.

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

Лексическая адаптация

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.

   a. Гав, говорю, идиотка!

   b. Я, конечно, обиделся.

Морфологическая адаптация

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**

Морфологическая адаптация - глагольные формы

Часть речи	**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

Синтаксическая адаптация - глагольные и именные группы

Часть речи	**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

Синтаксическая адаптация - связи внутри групп

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

Примеры адаптации

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.

Примеры адаптации

A. Едва оправясь от болезни, смотритель выпросил у С* почтмейстера отпуск на два месяца и, не сказав никому ни слова о своем намерении, пешком отправился за своею дочерью.

B. Как только смотритель почувствовал себя лучше, он попросил отпуск на два месяца и, не сказав никому ни слова о том, что он хочет делать, пошёл пешком за своей дочерью.

Статистическое тестирование

Коэффициент ранговой корреляции Кендалла между зависимой переменной и независимыми.

- Отрицательные корреляции: процент слов, входящих в лексические минимумы для ТРКИ, некоторые лексические списки, формула Оборневой;
- Положительные корреляции: 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).

Моделирование

Цель: исследование зависимости между признаками и классом текста.

Модель: логистическая регрессия, liblinear solver.

Средняя F1-мера: 70.5

Наиболее значимые признаки:

- В отнесении к классу адаптированных текстов: количество слов из лексических минимумов и некоторых списков лексики, фамилии, существительные;
- В отнесении к оригинальным: процент длинных слов, некоторые формулы удобочитаемости, сочинительные союзы, глубина групп

Выводы

В текстах RuAdapt особенно заметны следующие стратегии адаптации:

- Саммаризация;
- Удаление причастных и деепричастных оборотов;
- Уменьшение доли редких и “сложных” слов, замена их словами, которые должны быть знакомы читателям.

Результаты и следующие шаги

Параллельный датасет RuAdapt доступен здесь: https://github.com/Digital-Pushkin-Lab/RuAdapt. Данный датасет отличается от использованного в исследовании тем, что выровнен по параграфам. Его можно использовать для автоматического упрощения, а также для изучения стратегий адаптации.

Следующие шаги: исследование стратегий адаптации с привлечением экспертов-преподавателей РКИ; изучение потенциала собранного корпуса для автоматического упрощения текста.