**Приложение 1**

**Методика вычитки электронных текстов большого размера при помощи сервисов платформы www.corpus.by**

Ниже приведена методика вычитки электронного текста через программное обеспечение, разработанное сотрудниками Лаборатории распознавания и синтеза речи Объединённого института проблем информатики Национальной академии наук Беларуси. Исполнение данной методики позволяет получить вычитанный, орфографически правильный текст на белорусском языке.

Предложенным программным обеспечением являются сервисы обработки электронной текстовой информации, которые размещены на Интернет-платформе для обработки текста и речи www.corpus.by. Данное программное обеспечение работает в режиме онлайн и не требует установки на компьютер.

Суть методики вычитки состоит в обработке текста следующими онлайн-сервисами:

* «Подсчёт частотности символов»;
* «Проверка правописания»;
* «Проверка написания “Ў”»;
* «Идентификатор омографов».

По результатам обработки текста каждым сервисом пользователь может просматривать полученные результаты и по желанию вносить в текст соответствующие правки. Также результаты можно скопировать, вставить в файл \*.docx и сохранить.

Данная методика вычитки охватывает орфографический раздел правописания, но не охватывает пунктуационный и синтаксический разделы: правильность согласования слов и расстановка знаков препинания находится вне компетенции данной методики. Также методика опробована в многочисленных проектах Лаборатории распознавания и синтеза речи ОИПИ НАН Беларуси, поэтому рекомендовано осуществлять обработку сервисами в приведённом ниже порядке, который состоит из 6 последовательных этапов.

**Процесс вычитки**

Для получения орфографически правильного текста необходимо пройти 4 этапа вычитки через сервисы платформы [www.corpus.by](http://www.corpus.by) (рисунок 1). Перед началом внесения правок рекомендовано сделать копию проверяемого текста.



Рисунок 1. Главная страница компьютерно-лингвистической платформы [www.corpus.by](http://www.corpus.by)

На главной странице платформы для большего удобства работы нужно нажать кнопки: фильтр «Вычитка» и порядок «Логический» (рисунок 2). Останутся только сервисы, задействованные в вычитке текста.



Рисунок 2. Сервисы вычитки на платформе [www.corpus.by](http://www.corpus.by)

**Этап 1. Вычитка через сервис «Подсчёт частотности символов»**

Сервис «Подсчёт частотности символов» (рисунок 3) продемонстрирует список всех символов, использованных в тексте, и позволит выявить и исправить их ошибочное использование.



Рисунок 3.Интерфейс сервиса «Подсчёт частотности символов»

Для получения результатов необходимо скопировать и вставить текст в поле ввода, после чего нажать кнопку «Получить информацию о частотности символов!». Отобразятся результаты (рисунок 4).



Рисунок 4. Результаты работы сервиса «Подсчёт частотности символов»

Необходимо просмотреть итоговый список и проверить в нем следующие моменты:

* одинаковое ли количество скобок **( )**, **[]**;
* одинаковое ли количество двойных кавычек **“ ”** и **« »**;
* присутствуют ли в тексте одиночные кавычки **"**, которые не должны присутствовать, если уже используются двойные кавычки **“ ”** и **« »**;
* правильно ли используется дефис /-/, короткое /–/, длинное /—/ тире;
* присутствуют ли латинские буквы в кириллическом тексте.

Так, например, если количество левых и правых скобок не совпадает, то, скорее всего, в тексте есть пунктуационные ошибки.

По найденным ошибочным употреблениям символов необходимо внести правки в текст, перепроверить сервисом еще раз скорректированный текст, и перейти к следующему этапу вычитки.

Сервис «Подсчёт частотности символов» доступен по ссылке:

<https://corpus.by/CharacterFrequencyCounter/?lang=ru>

Подробная инструкция по использованию сервиса: <https://ssrlab.by/ru/3323>

**Этап 2. Вычитка через сервис «Проверка правописания»**

Сервис «Проверка правописания» (рисунок 5) выявляет:

***1. Слова, в которых использованы латинские символы.*** Если в словах кириллического написания есть визуально одинаковые латинские символы (/а/, /і/, /с/ и т.д.), то текст будет в дальнейшем повсюду обрабатываться неточно, поэтому такие символы нужно выявить и заменить кириллическими.

***2. Слова с ошибками.*** Слово, написанное с ошибкой, отсутствует в словарях и будет отмечено сервисом как неизвестное. В список неизвестных слов также попадают слова, которые не содержат ошибку, но отсутствуют в словаре. Также сервис дает возможность игнорировать определенные слова. Эта возможность может понадобиться при вычитке пользователем узкоспециального текста, чтобы исключить попадание заранее неизвестного сервису слова в список неизвестных и ускорить просмотр этого списка.



Рисунок 5. Интерфейс сервиса «Проверка правописания»

Чтобы получить список слов, в которых присутствуют ошибки или использованы латинские символы, нужно вставить текст в поле ввода и нажать кнопку «Проверить!». Неизвестные слова и слова с латинскими символами выводятся в виде списка с возможностью увидеть контекст их использования (рисунок 6).



Рисунок 6. Результаты работы сервиса «Проверка правописания»

Необходимо просмотреть список неизвестных сервису слов и найти среди них слова, написанные с ошибками. Можно, но необязательно, значком слева отметить правильно написанные слова и нажать «Перепроверить!» — список этих слов можно будет скопировать и вставить в поле исключений. Эта возможность полезна при работе с узкоспециальным текстом большого размера, где много терминов, имен собственных или др. По найденным ошибкам необходимо внести правки в текст, перепроверить сервисом еще раз скорректированный текст, и перейти к следующему этапу вычитки.

Сервис «Проверка правописания» доступен по ссылке:

<https://corpus.by/SpellChecker/?lang=ru>

Более подробная инструкция по использованию сервиса:

<https://ssrlab.by/ru/3334>

**Этап 3. Вычитка через сервис «Проверка написания “Ў”»**

Сервис «Проверка написания “Ў”» (рисунок 7) проверяет правильность использования в тексте букв «у» и «ў».

Алгоритм сервиса ищет символы /у/, /У/, /ў/, /Ў/ и смотрит на предыдущий символ, таким образом проверяя правильность использования.

Сервис имеет поле «Исключения», где помещаются актуальные на этот момент исключения из правил правописания буквы «ў». Поле можно редактировать: удалять введённые по умолчанию и добавлять необходимые пользователю исключения.



Рисунок 7. Интерфейс сервиса «Проверка написания “Ў”»

Для получения результатов необходимо скопировать и вставить текст в поле ввода, после чего нажать кнопку «Проверить!». Отобразятся результаты (рисунок 8).



Рисунок 8. Результаты работы сервиса «Проверка написания Ў»

Необходимо просмотреть список мест, где сервис нашел случаи неправильного употребления «у» и «ў». По найденным ошибкам нужно внести правки в текст, перепроверить сервисом еще раз скорректированный текст, и перейти к следующему этапу вычитки.

Сервис «Проверка написания “Ў”» доступен по ссылке:

<https://corpus.by/ShortUSpellChecker/?lang=ru>

Подробная инструкция по использованию сервиса:

<https://ssrlab.by/ru/1404>

**Этап 5. Вычитка через сервис «Идентификатор омографов»**

Сервис «Идентификатор омографов» (рисунок 9) показывает пользователю омографы – слова, которые имеют одинаковое написание и разное произношение, например, *му́зыка* і *музы́ка*.



Рисунок 9. Интерфейс сервиса «Идентификатор омографов»

Чтобы получить список использованных в тексте омографов, нужно вставить текст в поле ввода и нажать кнопку «Искать омографы!». За один раз сервис может обработать текст объемом около 20-30 страниц. Для удобства также поданы контексты, в которых находятся омографы (рисунок 10).



Рисунок 10. Результаты работы сервиса «Идентификатор омографов»

Необходимо внимательно просмотреть список найденных омографов и принять решение, в каких словах-омографах и на каких слогах в тексте нужно обозначить ударения, и внести эти ударения в текст.

Ударение в программе Microsoft Office Word добавляется следующим образом: нужно поставить курсор после буквы, на которой необходимо ударение, зажать клавишу Alt и набрать на цифровой клавиатуре справа последовательность цифр 0769.

Сервис «Идентификатор омографов» доступен по ссылке:

<https://corpus.by/HomographIdentifier/?lang=ru>

Более подробная инструкция по использованию сервиса:

<https://ssrlab.by/ru/4218>

Выполнение всех этапов данной методики позволяет получить вычитанный, орфографически правильный текст на белорусском языке.

Перечисленные в этой методичке сервисы и описанный алгоритм вычитки находится в состоянии постоянной доработки и усовершенствования. Лаборатория распознавания и синтеза речи приветствует все замечания и предложения по улучшению работы сервисов и данной методики.

**Контактные данные для обратной связи:**

Лаборатория распознавания и синтеза речи

Адрес: ул. Сурганова, 6, комнаты 422, 430 и 432

220012, г. Минск, Беларусь

Тел.: +375 (17) 284-27-73 (комната 422)

Факс: +375 17 284-21-75 (приёмная Института)

E-mail: yuras.hetsevich@newman.bas-net.by, ssrlab221@gmail.com

Национальная академия наук Беларуси

Объединенный институт проблем информатики