

1. ВВОД ТЕКСТА

Для набора текста в настоящее время широко используются две методики: ручной ввод и ввод с помощью «систем оптического распознавания» (OCR-систем). Для ввода текста Словаря Академии Российской были использованы обе методики. В результате выполнения этапа набора текста были сформированы текстовые файлы формата *.doc, при этом были определены следующие требования для текста словаря:

- сохранение разметки текста: текст набирается в 2 колонки, сохраняются начертания, размер символов;
- построчное соответствие текста оригиналу (в конце каждой строки – знак перехода строки);
- количество ошибок на странице не должно превышать 5-ти.

В процессе ввода текста САР осуществлялось создание специальной факсимильной шрифтовой гарнитуры, поэтому при вводе первых томов использовалась гарнитура Academia, последующие тома вводились с использованием уже созданной факсимильной гарнитуры Andrew Dashkova в соответствующих начертаниях. Для ввода текста шрифтом Academia ударные символы набирались с знаком апострофа «'», для специфических, устаревших символов использовались дополнительные знаки, например «#» – для «Ѣ», и т.п. После ввода текста первых томов необходимо было преобразовать его в новом шрифте Andrew Dashkova, для этого использовалась система замен в Word.

Технология ввода текста с помощью OCR-системы считается высоко эффективной в настоящее время, можно говорить, что в некоторых случаях ручной ввод текста занимает значительно больше времени, а следовательно менее эффективен. Эффективность OCR-систем зависит от множества факторов [Филиппович А.Ю., 2005-Ж]. Во-первых, правильные настройки системы, методика ввода, опыт оператора играет важную роль. Во-вторых, эффективность работы OCR-систем зависит от характеристик текстового оригинала: качества печати, разметки текста (верстки), шрифтовой гарнитуры, лексического состава и т.д. А также от характеристик самой OCR-системы: набор и состав функций, настроек, алгоритма распознавания и т.п. С учетом всех этих факторов тот или иной текст можно ввести за определенный промежуток времени. Для оценки эффективности использования OCR-систем для ввода текста САР подробно рассмотрена и исследована технология ввода.

1.1. Технология ввода текста с использованием системы оптического распознавания

Программы класса «системы оптического распознавания текстов»

Основное назначение систем оптического распознавания – это ввод текста для печатных и электронных изданий. Существует большое количество современных систем подобного рода за рубежом. Примерами могут служить следующие: *Recognita Plus DTK* фирмы Recognita Corporation (Венгрия), *TextBridge* фирмы Xerox Imaging Systems, *TypeReader* фирмы ExperVision (США), *CharacterEyes* фирмы Ligature (Израиль), *IRIS OCR* фирмы I.R.I.S. (Бельгия), *Easy Reader* фирмы Inovatic International (Франция), *OmniPage Professional* и *WordScan Plus* фирмы Caera (США) [Рынок OCR программ, 12.2005]. Среди программ, работающих под Linux, отметим *Clara OCR*, *Kadmos OCR/ICR*, *Vividata OCR Shop*, *GOCR*, *Ocrad*, *Kognition* [Обзор..., 12.2005]. Однако практически все они не поддерживают русский язык.

Наиболее известными OCR-системами в России являются: *OCR CuneiForm* [Cuneiform, 12.2005] и *ABBYY FineReader* [FineReader, 12.2005]. Основные возможности и функциональные характеристики последних версий ABBYY FineReader 8.0 Professional Edition и OCR CuneiForm 2000 Professional имеют схожий состав:

- современный интерфейс, панели быстрого доступа, мастер распознавания и сканирования, контекстная помощь, уроки работы в программе;
- сканирование с различных сканеров. Использование интерфейса TWAIN;
- импорт и обработка изображений различных форматов;
- автоматическая, ручная или полуавтоматическая, фрагментация изображений;
- распознавание полиграфических и машинописных гарнитур за исключением декоративных;
- возможность распознавания декоративных шрифтов с помощью обучения и создания эталонов;
- языковая поддержка;
- словарный контроль и возможность подключения и пополнения пользовательского словаря;
- поддержка WYSIWYG;
- распознавание и редактирование таблиц;
- интеграция с MS Word, MS Excel;
- пакетное сканирование и возможность организации распределенного параллельного сканирования в локальной сети.

Однако FineReader 8.0 – более современная версия и имеет более широкие возможности. Так, например, языковая поддержка включает 179 языков, для 36 предусмотрена проверка орфографии. Это связано с развитием другой линии продуктов компании ABBYY – электронными словарями Lingvo.

Кроме этого программа поддерживает больше форматов при импорте графических файлов (BMP, DCX, JPEG, PCX, PNG, TIFF, PDF) и экспорте. FineReader 8.0 позволяет экспортировать результаты распознавания в популярные офисные приложения, такие как Microsoft PowerPoint, Lotus Word Pro, Corel WordPerfect, Sun

StarWriter. Распознанный текст можно сохранить в следующих форматах: PDF, HTML, Microsoft Word XML, DOC, RTF, XLS, PPT, DBF, CSV, TXT и LIT.

Среди новых возможностей FineReader 8.0 отметим следующие:

- распознавание цифровых фотографий документов;
- дополнительные возможности при работе с PDF-файлами;
- автоматическая обработка документов;
- дополнительный режим для распознавания файлов с простым оформлением.

Согласно всем этим данным программа FineReader является более современной и обладает более широкими возможностями, поэтому она была выбрана для дальнейших исследований

Схема технологического процесса ввода текста с помощью системы оптического распознавания

Процесс ввода текста с помощью OCR-систем можно разделить на два этапа: предварительный и основной (рисунок 2.1). Первый включает в себя различные предварительные процедуры, общее назначение которых – настройка и подготовка инструментальных средств и рабочего места оператора для ввода текста. В общем случае этот этап может включать в себя следующие процедуры: установку и настройку аппаратных и программных средств, подготовку текста для ввода, настройку параметров системы оптического распознавания. Состав операций и процедур предварительного этапа зависит от уже существующих настроек системы.

Установка и настройка аппаратных средств может включать в себя следующие операции: установка, включение ЭВМ, подключение сканера к ЭВМ, установка драйверов и ПО для сканера и т.п. Для настройки параметров системы оптического распознавания необходимо проанализировать характеристики вводимого текста (качество оригинала, язык, лексику и т.д.) и, в зависимости от этого, настроить параметры сканирования и распознавания. Кроме этого для определения оптимальных настроек можно осуществить предварительный ввод небольшого объема текста. В этом случае следует проанализировать качества ввода и в зависимости от этого, изменять настройки системы оптического распознавания.

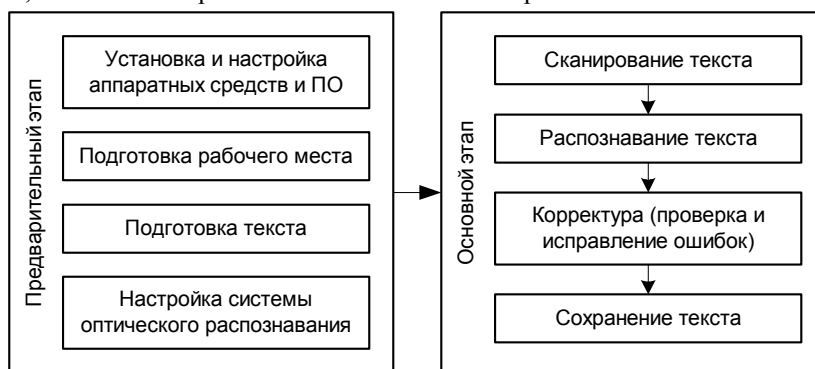


Рисунок 2.1. Обобщенная схема технологического процесса ввода текста с помощью системы оптического распознавания.

Второй, основной этап – это собственно ввод текста, он включает в себя четыре последовательные процедуры:

- сканирование;
- распознавание;
- корректурa, проверка и исправление ошибок;
- сохранение.

В соответствии с этой последовательностью организована работа в FineReader. Например, на панели Scan&Read расположены кнопки последовательности действий (рисунок. 2.2). Последовательность действий циклически повторяется для каждой страницы или ряда страниц.



Рисунок 2.2. Панель Scan&Read.

Однако, это наиболее общая методика ввода текста. В зависимости от характеристик исходного текста какие-то операции могут повторяться или быть исключены. Так, например, в некоторых случаях требуется повторное сканирование оригинала, если качество распознавания неудовлетворительно, и т.п.

Сканирование текста осуществляется с помощью специальной программы. Обычно она устанавливается с драйверами для сканера и специально предназначена для работы с определенной моделью сканера или целым модельным рядом. Например, для сканера Epson Perfection 2400 photo используется программа Epson Twain 5. Система оптического распознавания обращается к этой программе. Однако, сканирование можно также

осуществить в любом графическом редакторе. При этом выбирается опция импорта со сканера и также вызывается программа, используемая для сканирования (интерфейс Twain).

При сканировании страницы текста выполняется следующая последовательность операций:

- 1) установка страницы в сканер;
- 2) предварительное сканирование;
- 3) выбор сканируемой области;
- 4) анализ качества предварительного сканирования;
- 5) настройка параметров сканирования;
- 6) сканирование выбранной области (страницы) с заданными параметрами;
- 7) извлечение страницы из сканера.

Часто при сканировании все изображения страниц текста имеют схожие характеристики, поэтому настройка параметров сканирования требуется только вначале. Кроме этого, единство оформления и верстки издания в ряде случаев позволяют исключить операции 2-4.

После сканирования текста осуществляется его распознавание. При этом в FineReader возникает окно с изображением страницы, окно «текст» и окно укрупненного изображения, которые впоследствии будут использованы при корректуре.

В окне с изображением страницы следует выделить блоки для распознавания или использовать автоматическое выделение блоков. Процедура распознавания зависит от характеристик текста и его объема и выполняется автоматически.

При необходимости сохранения верстки для наилучшего результата рекомендуется вручную выделять и редактировать элементы для распознавания. Результат распознавания также зависит от настроек опций OCR-систем. Т.о. описывая процедуру распознавания можно выделить две операции: выделение блоков для распознавания и распознавание.

Проверка текста или корректура в большинстве случаев представляется наиболее трудоемкой и зависит от навыков оператора. После распознавания текста программа выделяет символы, форма которых вызвала сомнение при распознавании – *«неуверенно распознанные символы»*. Кроме этого, текст проверяется на орфографические ошибки с помощью словаря spellера, выявляя *«несловарные слова»*. Программа также позволяет откорректировать некоторые нарушения в наборе – пробелы после знаков препинания.

Процесс корректуры аналогичен проверке текста с помощью словаря spellера, используемого в текстовых редакторах. В Fine Reader появляется стандартное окно проверки (рисунок 2.4), в котором последовательно рассматриваются все помеченные символы и слова. Если встречается *«несловарное слово»*, то программа предлагает варианты исправления из слов словаря, отличных на один символ. Характеризуя процедуру корректуры в общем случае, можно выделить следующие операции: 1) сравнение проверяемого символа или слова; 2) исправление ошибки.

После проверки текста осуществляется его сохранение. При этом в FineReader предлагается передать текст в текстовый редактор (Microsoft Word) или другую программу. Данная процедура может только включать в себя операции по выбору параметров сохранения или передачи.

Порядок действий при распознавании текста в программе ABBY FineReader

1. **Запустить программу ABBY FineReader.**
2. **Создать новый пакет, сохранить его.**

Для этого необходимо использовать следующие команды: Файл → Новый пакет (Ctrl+N).

3. **Открыть отсканированное изображение страницы текста.** В качестве такого текста может выступать любой русский текст хорошего качества без рисунков, таблиц и схем.

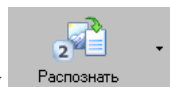
Для этого необходимо нажать либо кнопку



После этого возникает окно с изображением страницы, окно «текст» и окно укрупненного изображения, которые впоследствии будут использованы при корректуре.

4. **Распознать текст.**

Для распознавания текста необходимо нажать кнопку



После распознавания возникает окно «текст» и появляется сам распознанный текст.

При распознавании текста, имеющего специфическую верстку, для наилучшего результата рекомендуется вручную выделить и отредактировать элементы для распознавания. Для этого в окне «Изображение» необходимо выделить текстовые фрагменты, рисунки, таблицы с помощью соответствующих инструментов на панели слева.

Рекомендуется настроить опции распознавания, для этого в падающем меню «Распознать» необходимо выбрать опции, при этом появится следующее окно:

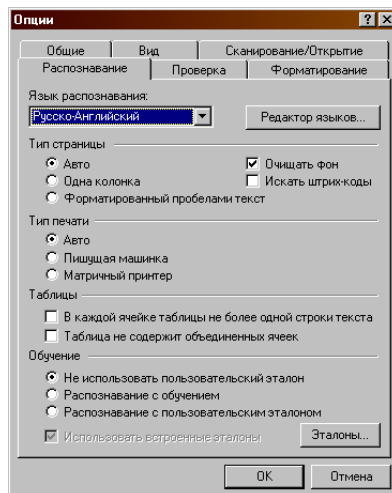
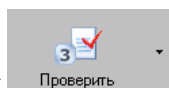


Рисунок 2.3. Окно «Опции» в Fine Reader.

5. Осуществить корректуру текста.



Для этого необходимо нажать кнопку

При этом появится окно проверки текста:

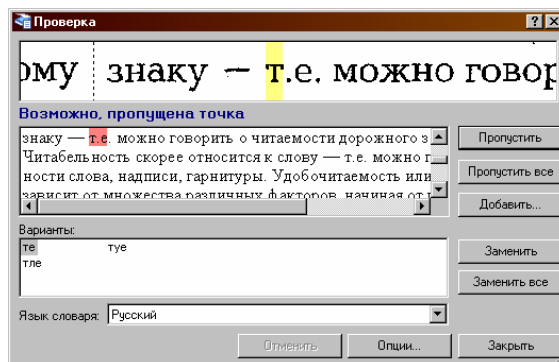
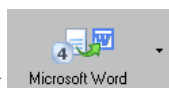


Рисунок 2.4. Окно проверки текста в Fine Reader.

Для наилучшего результата необходимо настроить опции проверки текста, для этого в падающем меню «Проверить» необходимо выбрать опции.

6. Передать полученный текст в Microsoft Word и сохранить полученный текстовый файл.



Для этого необходимо нажать кнопку

После этого будет запущена программа Microsoft Word и проверенный текст появится на экране. Его следует сохранить в формате *.doc, для этого в меню «Файл» необходимо выбрать «Сохранить».

Распознавание с обучением

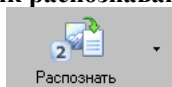
Программа ABBY FineReader обучена распознаванию стандартных шрифтов и не предназначена для распознавания декоративных шрифтов, например, FuturisXShadowC, PragmaticaShadowC, CyrillicGoth.

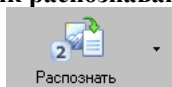
Для повышения качества распознавания данного документа воспользуемся специальным режимом распознавания: «распознавание с обучением». Обычно в данном режиме распознаются 1-2 страницы, в результате чего создается пользовательский эталон, который представляет собой набор изображений

символов, и в дальнейшем подключается для распознавания остальных страниц. При этом созданный эталон можно использовать только для распознавания текстов, использующих тот же шрифт и размер шрифта и отсканированных с тем же разрешением, что и документ, на основе которого данный эталон создавался.

Порядок действий при распознавании текста с обучением

1. **Выделить блоки на изображении** (меню Процесс→Анализ макета страницы).
2. **Установить режим «распознавание с обучением»** (на закладке Распознавание, меню Сервис→Опции в группе Распознавание с обучением установить переключатель в положение Распознавание с обучением). В строке состояния появится название эталона (по умолчанию default).
3. **Установить язык распознавания** (русский).



4. **Нажать кнопку** .
5. **Обучить эталон**, распознав страницу в режиме распознавания с обучением. Обучаемые символы заносятся в эталон, создаваемый системой по умолчанию. По окончании обучения система сохранит созданный эталон (по умолчанию в файл default.ptn) в папке, где хранится пакет.
6. **Отредактировать эталон.**

Далее необходимо отменить режим «распознавание с обучением» (на закладке Распознавание в группе Обучение установить переключатель в положение Распознавание с пользовательским эталоном).

Оценка эффективности работы системы оптического распознавания

Определяющим параметром эффективности работы OCR-системы является время, затрачиваемое на ввод текста: T при достаточном качестве ввода (количество ошибок на странице не должно превышать требуемого). Чтобы оценить временные затраты, проанализируем этапы технологического процесса ввода информации, подробно рассмотренного в предыдущем параграфе. Методика ввода текста включает два этапа: предварительный и основной. Предварительный этап технологического процесса зависит от текущих настроек системы. Так, если на предварительном этапе необходимо установить все аппаратное и программное обеспечение, требуемое для ввода текста, то временные затраты на это могут составить от нескольких часов до суток. В другом случае, если на предварительном этапе необходимо только запустить систему оптического распознавания и настроить опции распознавания и проверки, которые заранее определены, то на это потребуется несколько минут. Исходя из этого, предварительный этап технологического процесса ввода текста не рассматривается, тогда время ввода текста определяется следующим образом: $T = T_{осн. \text{ этапа}}$.

На основном этапе при вводе текста выполняются: сканирование, распознавание, корректура и сохранение текста.

$$T_{осн. \text{ этапа}} = T_{скан.} + T_{расп.} + T_{кор.} + T_{сохр.}$$

Данные операции выполняются для всех страниц текста.

$$T = \sum_{i=1}^m t_{скан_i} + \sum_{i=1}^m t_{расп_i} + \sum_{i=1}^m t_{кор_i} + \sum_{i=1}^m t_{сохр_i},$$

где m – количество страниц всего текста.

Рассмотрим подробнее процедуру корректуры текста. Будем учитывать только «неуверенно распознанные символы», т.к. проверка «несловарных слов» аналогична проверке при использовании автоматизированной методики корректуры и подробно рассмотрена в следующем параграфе.

В процессе проверки текста в FineReader осуществляется последовательная проверка всех слов, в которых встречаются неуверенно распознанные символы. При этом каждый такой символ помечен в тексте цветом. Если встречается ошибка, то осуществляется ее исправление и переход к следующему неуверенно распознанному символу, слову, где он встречается. Время проверки (корректуры) i -ой страниц определяется следующей формулой:

$$t_{кор_i} = n_{np_i} \cdot t_{cp} + n_{oi} \cdot t_u,$$

где t_{cp} – время сравнения проверяемого слова или символа, t_u – время исправления ошибки, n_{np_i} – количество проверяемых символов на i -ой странице, n_{oi} – количество ошибок на i -ой странице.

В результате получим, что для оценки эффективности набора текста с помощью OCR-систем следует осуществить сканирование (ввод) текста, распознавание, корректуру (исправление ошибок) и оценить временные затраты на выполнение этапов, процедур и отдельных операций технологического процесса ввода информации.

Наиболее важным параметром качества функционирования OCR-системы «является точность распознавания $A_{c_{расп}}$, обычно выражаемая процентным соотношением» [OCR&ICR Technology, 10.2006]. Для каждой i -ой страницы точность распознавания вычисляется:

$$A_{c_{расп_i}} = \frac{100\% \cdot n_{верно_расп_i}}{n_{общ_i}};$$

где $n_{верно_расп_i}$ и $n_{общ_i}$ есть количество верно распознанных символов и общее количество символов на странице соответственно.

$$n_{верно_расп_i} = n_{общ_i} - n_{о_i}, \text{ где } n_{о_i} - \text{ количество ошибок.}$$

Осуществляя корректуру текста, необходимо зафиксировать все проверяемые (неуверенно распознанные) символы и ошибки, допущенные при его вводе, посчитать количественные характеристики, провести анализ ошибок. В результате анализа ошибок должны быть выявлены типы ошибок, сделаны предположения о причинах их появления и описаны действия по их устранению.

1.2. Исследование эффективности использования OCR-систем для ввода текста CAP

Описание исследования

Для определения эффективности использования OCR-систем при вводе текста Словаря Академии Российской следует рассмотреть другие виды современных текстов хорошего и плохого качества. Исследование эффективности ввода текста с помощью систем оптического распознавания будет включать в себя следующие компоненты:

1. исследование временных затрат;
2. статистическое исследование количества ошибок;
3. анализ эффективности ввода текста.

При этом рассматриваются следующие виды текстов:

1. текст Словаря Академии Российской;
2. современный текст хорошего качества;
3. современный текст плохого качества.

Для проведения исследований текста Словаря Академии Российской были отобраны ксерокопии страниц разных томов с умеренным качеством изображений. Материалом для исследования современного текста хорошего качества послужили страницы современных книг, с хорошим качеством полиграфии, например [Галкин, 2003]. При исследовании современного текста плохого качества были подобраны фрагменты книг 50–80-х годов с плохим качеством печати, например: [Джермейн, 1973], [Чехов, 1969], [Гоголь, 1959], [Чехов, 1948], [Достоевский, 1979].

Исследование временных затрат

Материалы исследования временных затрат на ввод текстов представлены в приложении 2, таблицы П2.1, П2.2, П2.3. Случайно выбранные страницы были отсканированы и распознаны с помощью OCR-системы Abby FineReader 7.0. Итоговые средние значения для ввода страницы каждого из рассматриваемых текстов представлены в следующей таблице:

Таблица 2.1. Сравнение временных затрат на этапы ввода страницы текста.

Вид текста	Сканирование (сек.)	Распознавание (сек.)
текст Словаря Академии Российской	52	32
современный текст хорошего качества	62	29
современный текст плохого качества	61	30

Анализируя полученные данные, следует отметить, что время сканирования и распознавания зависит от множества факторов: характеристик сканера, производительности системы (скорости работы процессора, объема оперативной памяти и т.д.), от особенностей текста, качества оригинала, шрифта и т.п. Разработчики сканеров указывают скорость сканирования, производительность своих устройств, однако эти теоретические данные имеют приблизительный характер из-за множества факторов, влияющих на процесс сканирования: навыки оператора, эргономика его рабочего места, характеристики страниц (размера, качества и т.п.), опции сканирования и т.д. Некоторые исследования производительности сканирования (оцифровки изображений) представлены в статье [Филиппович А.Ю., 2001], а также в диссертации [Горбачев, 2006] и подтверждают полученные нами экспериментальные данные. Предполагается, что ввод текста будет осуществляться с помощью систем и программ, имеющих схожие характеристики, поэтому учет этих факторов не производился.

При сканировании страниц CAP использовался планшетный сканер Epson Perfection 2400 Photo. Были заданы следующие параметры: разрешение 300 dpi, размер страниц 17×22 см, черно-белое изображение

(GrayScale). При сканировании использовались ксерокопии страниц САР, сканирование с ксерокопий возможно произвести автоматически при использовании сканера с автоматической подачей бумаги. В этом случае время сканирования будет еще меньше, и это зависит от производительности сканера. При сканировании с оригинала (книги) возможно использование только планшетного сканера, при этом время сканирования будет большим, т.к. увеличатся затраты на подготовку оригинала (операции: открыть страницу, положить на сканер, прижать).

Временные затраты на сохранение страницы значительно меньше по сравнению с этапами сканирования и распознавания и составляют не более 1-2 секунды, также зависят от производительности системы.

Статистическое исследование количества ошибок

Согласно статистическим данным, полученным в результате исследования (приложение 2, таблица П2.4), среднее количество ошибок при вводе страницы текста САР составляет 286. Среднее количество символов на странице $Cp n_{Общ}=2065$. А усредненное значение точности распознавания будет равно: $Cp Ac_{расп}=86\%$.

В случае рассмотрения страниц современного текста хорошего или плохого качества (приложение 2, таблицы П2.5, П2.6) точность распознавания значительно выше и составляет соответственно для первого текста: $Cp Ac_{расп}=99,97\%$ ($n_{oi}=8$), для второго: $Cp Ac_{расп}=99,61\%$ ($n_{oi}=79$).

При вводе текста САР также использовалась функция распознавания с обучением, которая дает более качественный результат при вводе текстов нестандартных шрифтовых гарнитур (приложение 2, таблица П2.7). Точность распознавания значительно выше в этом случае и составляет: $Cp Ac_{расп}=95,16\%$ ($n_{oi}=103$).

Далее следует рассмотреть типы ошибок, которые встречались в тексте САР, и сделать предположения о причинах их возникновения.

Анализ типов ошибок, обнаруженный при вводе текста САР

Большая часть ошибок в тексте САР обусловлена особенностями графем шрифта и схожестью в написании символов. Все графемы символов алфавита шрифта состоят из элементов, единство форм которых обеспечивает единство рисунка всей шрифтовой гарнитуры. Фактически каждый типовой элемент повторяется в различных буквах алфавита, схожи формы овалов и полуовалов, основных и соединительных штрихов, всех тех элементов, что составляют основу рисунка графем символов. При распознавании это приводит к путанице символов. В таблице представлены примеры ошибок, связанных со схожестью графем символов в исследуемых текстах.

Таблица 2.2. Ошибки, связанные со сходством графем символов.

Ошибка	Пример ошибки в тексте	
	Распознанный текст	Исправленный текст
т → ш	Обетталость	Обетшалость
н → м	Сн.	См.
l → i	зляти	Зіяти
n → п, t → i,	пожранте	пожраніе
л → п, е → о, я → л,	лестыдятсл	постыдяться

В тексте САР используется старинная шрифтовая гарнитура, которая может быть отнесена к декоративному типу. Форма написания некоторых символов этой гарнитуры отличается от современных шрифтовых гарнитур. Например, схожими по рисунку в этой гарнитуре являются буквы: «ш» и «п». Фактически буква «п» является перевернутой буквой «ш», длинные засечки букв и соединительные штрихи приводят к тому, что эти символы при печати могут не отличаться друг от друга. Тем более в современных наборных шрифтах форма написания этой буквы другая, поэтому этот символ сложно правильно распознать.

Кроме этого в старинном тексте используются буквы, вышедшие из употребления, поэтому правильно распознать их невозможно. В основном в тексте [САР, 1789-1794] встречаются следующие старинные буквы: «ѣ», «ї», «ѵ», «ѿ». Очень большое количество ошибок связаны с неправильным распознаванием букв «ѣ» и «ѵ».

Следующая группа ошибок – ошибки, связанные с качеством оригинала. Таких ошибок сравнительно немного в современном тексте хорошего качества. Из-за загрязненности сканируемого изображения и пыли в распознанном тексте появились лишние знаки препинания («.», «.», «”»). При наличии карандашных помет в тексте неправильно распознаются символы, некоторые примеры представлены в таблице 2.3:

Таблица 2.3. Ошибки, связанные с качеством оригинала.

Ошибка	Пример ошибки в тексте	
	Распознанный текст	Исправленный текст
' → -	^^e^^^ми	стей пасти
« → и	Раз^воре^^e ^ел^^	- Разтвореніе челю-
? → даю	« ?»	и даю

Несколько ошибок вызваны неправильным выделением символов. Так два символа распознаны как один. Эти ошибки связаны с качеством печати или с небольшим размером межбуквенного интервала.

Таблица 2.4. Ошибки, связанные с некорректным выделением символов.

Ошибка	Пример ошибки в тексте	
	Распознанный текст	Исправленный текст
м → сл	мая	злая
зн → ш	зная	шая
М → че	ЛозолоМный	Позолоченый

Последняя группа ошибок – это нарушения технических правил в наборе: изменение начертания символов, отсутствие переносов, путаница в строчных и прописных символах:

- НаЗВаНие → название;
- ДверИ → двери;
- непо нятного → непонятного.

Наибольшее количество ошибок такого рода (≈50 ошибок на странице) – это замена строчной буквы «Ъ» на конце согласных прописной буквой, которая происходит из-за похожей высоты сточного и прописного твердого знака.

На основе исследования типов ошибок была составлена таблица типовых ошибок распознавания и разработана методика устранения некоторых типовых ошибок. В этом случае использовалась система замен.

Качество распознавания зависит от двух факторов: от эффективности математических методов оптического распознавания символов и эффективности лингвистической компоненты, используемой при распознавании. При этом в состав лингвистической компоненты входят различные словари и грамматика ([Организация взаимодействия, 1990], [Бахмутский, 1986], [Бабко-Малая, 1987], [Воскресенский, 1989], [Автоматизация..., 1989]). Для улучшения качества распознавания была использована словарная компонента. Для этого был составлен словарь-спеллер, который содержал основные словоформы текста САР. При подключении словаря и при использовании системы замен типовых ошибок распознавания (приложение 2, таблица П2.8) среднее количество ошибок на странице составило $n_{oi}=53$, точность распознавания при этом: $Cp Ac_{расп} = 98,93\%$. В случае использования словаря-спеллера примерно в два раза уменьшается количество проверяемых символов.

Результаты исследования

Для оценки эффективности систем оптического распознавания воспользуемся схемой, представленной ранее. Время ввода текста с помощью OCR-системы определяется следующим выражением:

$$T = \sum_{i=1}^m t_{скан_i} + \sum_{i=1}^m t_{расп_i} + \sum_{i=1}^m t_{кор_i} + \sum_{i=1}^m t_{сохр_i}$$

Используя статистические данные, полученные в результате исследования, осуществим оценку временных затрат на ввод страницы текста. Время ввода *i*-ой страницы определяется следующим выражением:

$$t_i = t_{скан_i} + t_{расп_i} + t_{кор_i} + t_{сохр_i}$$

Согласно статистическим данным исследования различных видов текстов (*1 текст* – текст САР, *2 текст* – современный текст хорошего качества, *3 текст* – современный текст плохого качества) временные затраты на следующие этапы ввода текста – сканирования, распознавания и сохранения – равны:

$$t_i^{1\text{текст}} = t_{скан_i}^{1\text{текст}} + t_{расп_i}^{1\text{текст}} + t_{кор_i}^{1\text{текст}} + t_{сохр_i}^{1\text{текст}} = 52 + 32 + t_{кор_i}^{1\text{текст}} + 1 = 85 + t_{кор_i}^{1\text{текст}}$$

$$t_i^{2\text{текст}} = t_{скан_i}^{2\text{текст}} + t_{расп_i}^{2\text{текст}} + t_{кор_i}^{2\text{текст}} + t_{сохр_i}^{2\text{текст}} = 62 + 32 + t_{кор_i}^{2\text{текст}} + 1 = 95 + t_{кор_i}^{2\text{текст}}$$

$$t_i^{3\text{текст}} = t_{скан_i}^{3\text{текст}} + t_{расп_i}^{3\text{текст}} + t_{кор_i}^{3\text{текст}} + t_{сохр_i}^{3\text{текст}} = 61 + 30 + t_{кор_i}^{3\text{текст}} + 1 = 92 + t_{кор_i}^{3\text{текст}}$$

Время ввода зависит от времени, затрачиваемом на корректуру текста. Временные затраты на остальные этапы отличаются не более чем на 10 секунд, что является допустимой погрешностью. Тогда при сравнении временных затрат на ввод страницы для современного текста хорошего качества, современного текста плохого качества и старинного текста будет учитываться только время корректуры.

$$t_{кор_i} = n_{np_i} \cdot t_{cp} + n_{o_i} \cdot t_u,$$

где t_{cp} – время сравнения проверяемого символа, t_u – время исправления ошибки, n_{np} – количество проверяемых символов на странице, n_o – количество ошибок на странице.

Пусть $t_{cp}=t$, $t_u=Kt$, тогда $t_{кор} = n_{np} \cdot t + n_o \cdot Kt$

Согласно статистическим данным исследования, количество проверяемых символов для рассматриваемых текстов будет равно:

$$n_{np_i}^{1\text{текст}} = 428, n_{np_i}^{2\text{текст}} = 102, n_{np_i}^{3\text{текст}} = 234$$

А среднее количество ошибок в рассматриваемых текстах:

$$n_{o_i}^{1\text{текст}} = 286, n_{o_i}^{2\text{текст}} = 8, n_{o_i}^{3\text{текст}} = 79$$

В случае использования для проверки эталона и словаря spellera и системы замен количество ошибок на странице CAP равно:

$$n_{np_i}^{1\text{текст}} = 197, n_{o_i}^{1\text{текст}} = 53$$

Тогда время корректуры:

$$t_{кор_i}^{1\text{текст}} = n_{np_i}^{1\text{текст}} \cdot t + n_{o_i}^{1\text{текст}} \cdot Kt = 428t + 286Kt$$

$$t_{кор_i}^{2\text{текст}} = n_{np_i}^{2\text{текст}} \cdot t + n_{o_i}^{2\text{текст}} \cdot Kt = 102t + 8Kt$$

$$t_{кор_i}^{3\text{текст}} = n_{np_i}^{3\text{текст}} \cdot t + n_{o_i}^{3\text{текст}} \cdot Kt = 234t + 79Kt$$

В случае использования для проверки эталона, словаря spellera и системы замен необходимо добавить время на обучение эталона, тогда время корректуры равно:

$$t_{кор_i}^{1\text{текст}} = t_{обучения} + (n_{np_i}^{1\text{текст}} \cdot t + n_{o_i}^{1\text{текст}} \cdot Kt) = t_{обучения} + (197t + 53Kt)$$

Если коэффициент $K=1$, т.е. время проверки символа и время исправления ошибки равны, то

$$t_{кор_i}^{1\text{текст}} = 428t + 286t = 714t$$

$$t_{кор_i}^{2\text{текст}} = 102t + 8t = 110t$$

$$t_{кор_i}^{3\text{текст}} = 234t + 79t = 313t$$

Если коэффициент $K=10$, то

$$t_{кор_i}^{1\text{текст}} = 428t + 2860t = 3288t$$

$$t_{кор_i}^{2\text{текст}} = 102t + 80t = 182t$$

$$t_{кор_i}^{3\text{текст}} = 234t + 790t = 1024t$$

Без учета времени на обучение эталона в случае использования для проверки эталона, словаря spellera и системы замен время корректуры:

$$t_{кор_i}^{1\text{текст}} = 197t + 53Kt.$$

$$\text{При } K=1: t_{кор_i}^{1\text{текст}} = 197t + 53t = 250t,$$

$$\text{при } K=10: t_{кор_i}^{1\text{текст}} = 197t + 530t = 727t.$$

Сравним полученные показатели. В случае если время сравнения символа и исправления равны ($K=1$), корректура одной страницы Словаря Академии Российской займет 6,5 раз больше времени по сравнению с корректурой современного текста хорошего качества, в 2,3 раза больше по сравнению с современным текстом плохого качества. Если время исправления ошибки в 10 раз больше времени сравнения ($K=10$), то корректура одной страницы Словаря Академии Российской займет 18 раз больше времени по сравнению с корректурой современного текста хорошего качества, в 3,2 раза больше по сравнению с современным текстом плохого качества.

В случае использования для проверки текста CAP эталона, словаря spellera и системы замен время корректуры текста CAP в 2,3 (при $K=1$), в 4 (при $K=10$) раза больше по сравнению с текстом хорошего качества; в 0,8 (при $K=1$), в 0,7 раз меньше по сравнению с текстом плохого качества.

Сравним полученные значения точности распознавания текстов:

$$Ac_{расп}^{1\text{текст}} = 86\%, Ac_{расп}^{2\text{текст}} = 99,97\%, Ac_{расп}^{3\text{текст}} = 99,61\%.$$

В случае использования для проверки текста CAP эталона, словаря spellera и системы замен:

$$Ac_{расп}^{1\text{текст}} = 98,93\%.$$

Низкое значение точности распознавания (86%) не позволяет рекомендовать использование OCR-системы для ввода текста CAP, однако использование эталона, словаря-spellera и системы замен для типовых ошибок распознавания дает значительно лучший результат $\approx 99\%$, что соизмеримо с данными, полученными при вводе текста плохого качества, в этом случае следует использовать данную методику для ввода.

2. КОРРЕКТУРА ТЕКСТА

2.1. Исследование эффективности корректурных процессов

По завершении этапа набора текста осуществляется корректура. Обычно она состоит из нескольких чтиток в зависимости от типа, назначения издания и качества набора [Зелинская, 2002, с. 9].

В «традиционном классическом описании» корректура состоит из двух основных процессов: чтения корректурных оттисков и правки набора. Оттиски с набора читают корректоры, сличающие их с оригиналом или с предыдущими корректурными оттисками, а также авторы и редакторы, проверяющие правильность введенного текста по существу. При чтении оттисков ошибки отмечаются специальными корректурными знаками, повторяемыми на полях оттисков, причем рядом с этими знаками указываются правильные буквы, слова и т.п. [Гулько, 1995]. После этого правка с корректурных оттисков вносится в набор. Правила и рекомендации корректуры различных типов изданий ориентированы на современное представление о верстке текста и представлены в различных справочных пособиях, например [Справ. пособие, 1984], [Справ. книга, 1985], [Рисс, 1980], [Рыжова, 2005].

В современной технологии допечатного процесса на основе средств вычислительной техники корректура выполняется в текстовых процессорах и программах верстки. В связи с этим понятия, описывающие корректуру, изменились. Так, говоря о наборе, подразумевают ввод и формирование электронного документа, а под оттиском набора – распечатку этого документа. Кроме этого современные текстовые процессоры содержат встроенные функции проверки текста на наличие грамматических, синтаксических и стилистических ошибок. Одна из них – функция *спеллер* (speller – сокращение от spelling checker – программа поиска опечаток, корректор [Борковский, 1989]) позволяет автоматизировать корректуру и редактирование текста, снизить временные затраты прежде всего на поиск ошибок в написании слов.

Процесс корректуры регламентирован лишь в основном, и на его конкретное содержание и результаты оказывают влияние множество различных факторов:

- во-первых, особенности издания (первое издание или какое-либо его переиздание);
- во-вторых, индивидуальные особенности текста (тема, предмет, язык, авторские цели, назначение и т.п.);
- в-третьих, профессионализм корректора (культурный уровень, знания, навыки, умения, психологические установки, социально-экономические факторы и др.);
- в-четвертых, технологические факторы (форма рабочего материала, инструментальные аппаратные и программные средства поддержки корректорской деятельности, временные и стоимостные ресурсные ограничения, методика и др.).

Некоторые теоретические компоненты, обобщенные представления устройства программ проверки текста представлены в различных источниках: [Автоматизация..., 1989], [Бабко-Малая, 1987], [Бахмутский, 1986], [ИТ ввода..., 1998], [Языковые средства..., 1990], [Воскресенский, 1989], однако математические алгоритмы работы конкретных программных продуктов держаться в секрете. В литературе в основном представлены правила и рекомендации корректуры различных типов изданий, и не представлены аналитические модели технологий корректуры. Особенностью современных программ проверки текстов является их ориентация на современную общеупотребительную лексику, что затрудняет их использование для специфических, старинных текстов. Эффективность использования различных программ проверки для текстов со специфической лексикой малоисследована.

Цель настоящего исследования корректурных процессов – это выявление возможности автоматизации корректуры с использованием спеллеров и определение ее эффективности. При этом решались следующие задачи: 1) построение формальных моделей традиционной и автоматизированной методик корректуры; 2) статистические исследования ошибок в тексте Словаря; 3) частотные исследования текста Словаря на малой и большой выборке.

Формальные модели методик корректуры

Регламентация корректурных процессов носит в основном общий характер, прежде всего, из-за индивидуальных особенностей текстов и разнообразия собственных методик, которые используют корректоры. Во всех случаях в инструментарий корректора обязательно входят различные словари. Современная форма словарей – это не только последние печатные издания, но и различные электронные лексикографические ресурсы, в числе которых электронные словари на CD ROM, Интернет-порталы, словарные базы данных, встроенные в текстовые редакторы и издательские системы орфо- и грамматические редакторы, программы спеллеры и т.п. Электронные ресурсы рассматриваются как современные средства автоматизации корректорской и редакторской деятельности, однако величина эффекта от их использования может оказаться незначительной или вовсе отсутствовать.

Например, в одних случаях удобно использовать спеллер, в других – нет. Спеллер обычно содержит наиболее часто употребляемые слова. Если мы имеем дело с текстами, содержащими специфическую лексику, то количество «ошибок», найденных автоматически, будет достаточно велико. Этими «ошибками» назовем слова, которые отсутствуют в словаре спеллера. Поэтому тексты со специфической лексикой проверяют корректоры, хорошо знающие эту лексику, а также используя предметные и терминологические словари. Корректор опирается на собственный опыт и, имея дело с определенным текстом, быстро находит типовые ошибки.

Рассмотрим две методики корректуры, условно названные нами «традиционной» и «автоматизированной». «Автоматизированная» методика отличается от «традиционной» тем, что в ней используется спеллер с функцией пополнения словаря [Филиппович А.Ю., 2005-В]. Оценим эффективность этих методик путем исследования их формальных моделей (рисунки 2.5, 2.6).

Традиционная методика корректуры

Корректор проверяет текст последовательно страницу за страницей. Он сравнивает пословно текст с его оригиналом. Время, затрачиваемое на корректуру, определяет эффективность его работы. Обозначим время корректуры i -ой страницы текста как $t_{\kappa i}$. Оно будет определяться через следующее выражение:

$$t_{\kappa i} = n_i \cdot t_{cp} + n_{oi} \cdot t_u,$$

где: t_{cp} – время сравнения слова, t_u – время исправления ошибки

n_i – общее количество слов на i -ой странице, n_{oi} – количество ошибок на i -ой странице.

Соответственно время корректуры всего текста определяется следующим выражением:

$$T_k^t = \sum_{i=1}^m t_{\kappa i} = \sum_{i=1}^m n_i \cdot t_{cp} + \sum_{i=1}^m n_{oi} \cdot t_u,$$

где m – количество страниц всего текста.

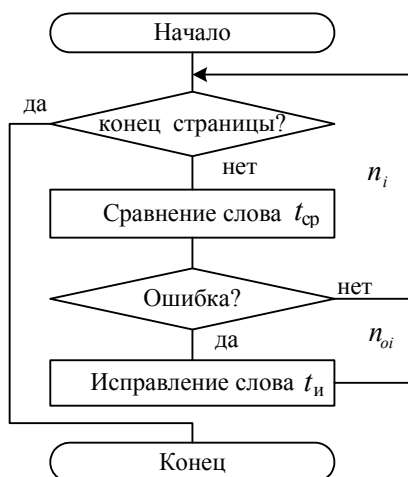


Рисунок 2.5. Традиционная методика корректуры страницы текста.

Анализируя модель данной методики, можно отметить, что здесь фигурируют два типа параметров: время, затрачиваемое на ту или иную деятельность корректора, и количество слов.

Время сравнения слова с оригиналом и время исправления слова определяются профессионализмом корректора, его квалификацией. Другими параметрами, от которых зависит эффективность корректуры, является количество слов, просматриваемых корректором – n_i , и количество ошибок на странице n_{oi} . Изменение этих параметров позволяет влиять на эффективность процесса корректуры.

Автоматизированная методика корректуры

Методика корректуры с использованием спеллера позволяет автоматизировать процесс проверки ошибок. Корректор последовательно проверяет страницу за страницей текста. Однако он проверяет не все слова, а только слова, неизвестные компьютеру. Эти слова помечены, например, в Word они подчеркнуты волнистой цветной (красной) линией. Каждое правильное неизвестное слово после проверки заносится в словарь. Т.о. по мере пополнения словаря количество неизвестных слов уменьшается на каждой последующей странице.

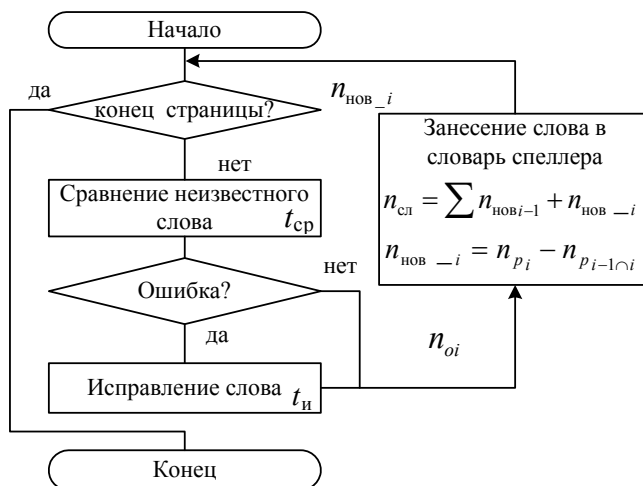


Рисунок 2.6. Автоматизированная методика корректуры страницы текста.

Предположим, что словарь spellера пустой, тогда все слова первой страницы будут «новыми» – неизвестными. На каждой последующей странице слова будут делиться на те, которые уже встречались – «старые», и те, которые не встречались ранее – «новые».

Тогда время проверки страницы определяется следующей формулой:

$$1\text{-ая страница: } t_{\kappa 1} = n_{нов_1} \cdot t_{cp} + n_{o1} \cdot t_u,$$

где $n_{нов1}$ – количество новых слов на первой странице, n_{oi} – количество ошибок на первой странице.

Количество новых слов – занесенных в словарь: $n_{сл} = n_{нов_i} = n_{p_i}$, где n_{p_i} – количество разных слов на i -ой странице (неповторяющихся на странице).

$$2\text{-ая страница: } t_{\kappa 2} = n_{нов_2} \cdot t_{cp} + n_{o2} \cdot t_u,$$

$$n_{нов_2} = n_{нов_1} + n_{нов_2}, \quad n_{нов_2} = n_{p_2} - n_{p_1 \cap 2},$$

где $n_{p_1 \cap 2}$ – количество общих разных слов 1-ой и 2-ой страниц.

...

$$i\text{-ая страница: } t_{\kappa i} = n_{нов_i} \cdot t_{cp} + n_{oi} \cdot t_u,$$

$$n_{сл} = \sum n_{нов_{i-1}} + n_{нов_i}, \quad n_{нов_i} = n_{p_i} - n_{p_{i-1} \cap i}$$

2.2. Анализ ошибок корректуры

Исследование количества ошибок

Цель данного исследования – это выявить среднее количество ошибок в тексте Словаря для того, чтобы оценить параметр n_{oi} (количество ошибок на i -ой странице).

В качестве источника исследования был взят фрагмент текста 1-го тома САР – раздел «Показание». Этот раздел представляет собой указатель слов словаря. Для сравнения были взят текст Показания, полученный при вводе текста, и итоговый вычитанный вариант. Тексты были обработаны в Word с помощью замен для последующего импорта в таблицы БД. В результате была сформирована таблица текста, полученного при вводе – Pok1tOsh, и таблица вычитанного текста – Pok1t. Далее с помощью запросов были выявлены количественные характеристики таблиц. Таблицы имеют следующую структуру: $\langle W_z, K \rangle$, где W_z – слово (словосочетание) Показания, K – номер колонки. Сравнивая таблицы Pok1tOsh и Pok1t мы получили таблицу ошибок (приложение 2, таблица П2.9).

Таблица 2.5. Результаты сравнения ошибок в «Показании» САР 1-го тома

Характеристики сравнения (Количество)	Введенный текст (табл. Pok1tOsh)	Вычитанный текст (табл. Pok1t)
Всего записей	6092	6103
Всего неповторяющихся записей	6078	6094
Всего слов	6092	6103
Всего неповторяющихся слов	6031	6049
Одинаковых записей	5499	
Одинаковых неповторяющихся записей	5477	
Одинаковых слов	5731	
Одинаковых неповторяющихся	5571	

слов		
Ошибок в неповторяющихся записях	601	
Ошибок в неповторяющихся словах	460	
Ошибок в номерах колонок	108	
Отсутствующих записей	11	
Отсутствующих слов	11	
Отсутствующих номеров колонок	33	

Общее количество несоответствий (ошибок) в тексте Показания составляет 612 ошибок. Общий объем текста Показания составляет 46 страниц. Таким образом, среднее количество ошибок на странице составляет 13,3. Если считать, что ошибки распределены равномерно по всему тексту словаря, тогда на одной странице будет встречаться 13-14 ошибок.

Анализ систематических ошибок

В результате исследования ошибок в тексте раздела «Показание» первого тома САР были выявлены некоторые систематические ошибки.

Таблица 2.6. Ошибки, связанные со старинной лексикой и грамматикой.

Описание ошибки	Примеры		Кол-во ошибок
	Ошибки	Исправления	
Отсутствие Ъ на конце	Абшип	Абшипѣ	17
	Бекеп	Бекепѣ	
ѣе→ѣ	Вдовец	Вдовецѣ	20
	Бѣлоручка	Бѣлоручка	
	Бѣщуся	Бѣщуся	
иі→ї, ий→їй	Набѣгѣ	Набѣгѣ	16
	Повязывание	Повязываніе	
ие→їе	Вороній	Вороній	16
	Провѣщаніе	Провѣщаніе	
Старинное написание слов	Бальсамический	Бальсамическій	≈20
	Воинский	Воинскій	
	Орудие	Орудіе	
	Збывание	Збываніе	
	Бадьянѣ Сибирский	Бадьянѣ Сибирской	
	Вельможеспво	Вельможспво	
	Испровергаются	Испровергаюся	
Оружебормый	Оружебормый		
Подбираюсь	Подбиваюся		
Ублаженіе	Ублажаніе		

Часть ошибок связана со старинной славянской лексикой и грамматикой (27%), используемой в Словаре (таблица 2.6). Это касается и специфических символов – букв старого алфавита, и целых слов, которые уже вышли из употребления. Ярким примером подобной ошибки является отсутствие «ѣ» после согласных на конце слов. Множество ошибок этого типа связаны с символами, которые не входят в современный алфавит:

«Ѣ» (ять) и «ї». В современном языке вместо этих букв употребляются буквы «е» и «и». Чаще всего в тексте Показания встречается сочетания «Ѣе» и «иї», что может быть связано с какой-то систематической ошибкой замены или с особенностями ввода текста.

Ошибки в написании старинных слов характерны для некоторых окончаний: Сибирской (Сибирский), Подбиваюся (Подбираюсь) и суффиксов: Вельможство (Вельможество), Испровергаюся (Испровергаюсь), Ублажаніе (Ублаженіе).

Другая группа ошибок обусловлена особенностями графем шрифта (21%), используемого при наборе, и схожестью в написании символов. Похожий рисунок графем некоторых символов приводит к их путанице.

Примером могут служить буквы «ш» и «п», «щ» и «щц». Клавиши с буквами «ш» и «щц» находятся рядом, поэтому данная ошибка может быть вызвана этим. Примером похожих букв с округлыми элементами являются «ь», «ъ» и «ѣ».

Таблица 2.7. Ошибки, обусловленные особенностями графем шрифта.

Символы	Ошибки	Примеры Исправления	Кол-во ошибок
п ← → ш	АскишѢ	АскипѢ	13
	Баронсшво	Баронспво	
	Волишель	Волипель	
	ВоропникѢ	ВорошникѢ	
	Наблопняю	Налошняю	
	Обеппалоспѣ Блудяшіе огни	Обепшалоспѣ Блудяшіе огни	
ш ← → щ	Вѣщаніе	Вѣшаніе	3
	Всевысочайще АппекаревѢ	Всевысочайше АппекаревѢ	
	БарвенокѢ	БарвенокѢ	
	Безѣизбѣжно	Безѣизбѣжно	
ѣ ← → ѣ	Внѣ	Внѣ	32
	Вывѣвки	Вывѣвки	
	Единовѣрїе БилѣрдѢ	Единовѣрїе БилѣрдѢ	
	Вѣспѣ	Вѣспѣ	
ѣ ← → ѣ	НеворопѢ АспидовѢ	Неворопѣ Аспидовѣ	5
	Вѣспѣ	Вѣспѣ	
	Вѣспѣ	Вѣспѣ	
л ← → д	Водохранидище	Водохранилище	3
	Водохранилище	Водохранилище	

Согласно исследованию, наибольшее количество ошибок было связано с путаницей «ѣ» и «ѣ» (таблица 2.7), данные буквы являются достаточно сложными для набора и с точки зрения старинного словоупотребления, и с точки зрения особенностей графем. Сравнительно редко появляется ошибка из-за схожести графем букв «л» и «д», таких ошибок было выявлено только 3. В результате анализа также были выявлены другие одиночные ошибки, которые могут быть связаны с особенностями написания, например в буквах «л» и «п»: ЛирѢ – пирѢ; «е» и «с»: Верепѣ – Верспѣ; «л» и «я»: Выл – Выя; «ѣ» и «б»: Оьладапель – Обладапель.

Другие систематические ошибки составляют около 51%. Среди них технические ошибки набора (таблица 2.8). Эти ошибки связаны с характеристиками программ и технических средств, используемых для ввода

текста. В Показании часто вместо точки встречается буква «ю». Причиной этому может быть близкое расположение этих знаков на клавиатуре. Но скорее всего эта ошибка вызвана использованием латинского и русского регистров. Дело в том, что на латинице «точка» расположена на букве «ю» кириллицы.

Другим примером технической ошибки является наличие прописных букв после точки. Это связано с настройками текстового редактора. Так, в Microsoft Word при наборе текста система автоматически после точки через пробел ставит прописную букву. В Показании для некоторых повторяющихся слов указывается часть речи или краткое пояснение. В качестве разделителя в этом случае используется точка: Вьюрокъ. Ппашка.

Таблица 2.8. Технические ошибки.

Описание ошибки	Примеры		Кол-во ошибок
	Ошибки	Исправления	
	АзѢ мѣспою	АзѢ мѣспои.	
ю→.	Балакирю	Балакирь.	12
Прописные буквы после точки	Изневѣспью Аа. Межд	Изневѣспь. Аа. Межд	
	Альпѣ. Скрыпка	Альпѣ. Скрыпка	35
..Н←→.н	Вьюрокъ. Ппашка	Вьюрокъ. Ппашка	

В тексте Показания обнаружены систематические ошибки, которые трудно отнести к какой-либо группе и определить их причины (таблица 2.9). Самое большое количество ошибок (49) это наличие пробелов в словах, особенно перед символом «ѣ». Другой пример – это ошибки-опечатки: повторение символов и сочетаний а также их перестановка.

Таблица 2.9. Другие систематические ошибки.

Описание ошибки	Примеры		Кол-во ошибок
	Ошибки	Исправления	
	Б ѣца	Бѣца	
ıı, ııe, ııй, ıı_e, ıı_й	Баснослов ѣе	Баснословіе	
	Бомбардирск ѣй	Бомбардирскѣй	49
	Благосовѣп ѣ е	Благосовѣпѣе	
і→ї	Боярск і й Визжаніе	Боярскѣй Визжаніе	
	Забѣганіе	Забѣганіе	19
Повторение символов	Завоеваніе Арканѣѣ	Завоеваніе Арканѣ	
	Брезгунька	Брезгунька	25
Перестановка символов	БѣБѣлена Бабки вольчи	БѣБѣлена Бабки волчѣи	
	Сповѣлеваю	Сповѣлѣваю	2

Характер некоторых систематических ошибок свидетельствует о том, что данный текст был сформирован с помощью ручного набора. На это указывают некоторые технические ошибки и опечатки, которые мог сделать только человек.

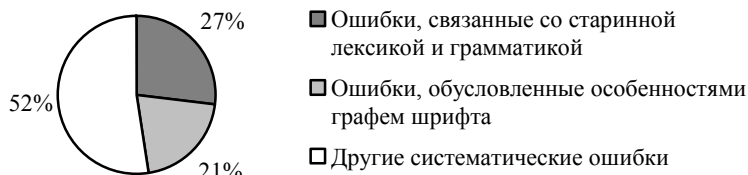


Рисунок 2.7. Соотношение разных

видов ошибок.

Некоторые систематические ошибки можно устранить автоматически с помощью замен, что уменьшит временные затраты на корректуру. Для устранения ошибок, связанных со старинной лексикой и грамматикой рекомендуется проверить и заменить следующие окончания и суффиксы: ий→ий, ие→ие, юпся→юся. Для технических и других систематических ошибок: і→ї, иї→ї, ъе→ъ, _ї→ї.

2.3. Исследование частотных характеристик слов

Исследование на малой выборке

Данное исследование проводится с целью определить характер изменения количества новых слов на каждой последующей странице текста.

Исследование проводится на малой выборке текста. Суть исследования состоит в следующем. Рассматриваются 8 первых страниц САР 1-го тома. Каждая последующая страница сравнивается с предыдущими: вторая с первой, третья с первой и второй и т.д. В результате сравнения необходимо определить количественные характеристики слов: общее количество слов на странице, количество разных слов, количество слов, которые встречались ранее и, соответственно, количество новых слов, также рассматриваются слова, известные и не известные компьютеру (входящие и не входящие в словарь spellera).

Таблица 2.10. Характеристики страниц 1-8.

Характеристики сравнения	Страницы							
	1	2	3	4	5	6	7	8
Общее количество слов на странице	228	256	279	268	265	294	276	288
Общее количество слов, известных Word	108	112	134	135	134			
Общее количество слов, не известных Word	120	144	145	133	131			
Количество разных слов	188	201	227	211	215	233	222	226
Количество разных слов, известных Word	88	91	101	99	103	115	113	113
Количество разных не известных Word	100	110	126	112	112	118	109	113
Общее количество ранее встречавшихся на странице слов		51	86	103	101	130	107	125
Количество разных слов, ранее встречавшихся на странице		24	41	58	56	83	69	78
Количество разных слов, ранее встречавшихся и известных Word		13	26	32	34	46	38	51
Количество разных слов, ранее встречавшихся и не известных Word		11	15	26	22	37	31	27

Последовательность проведения исследования следующая. Сначала были взяты тексты первых восьми страниц Словаря. Далее они были обработаны в Word: удалены все знаки препинания, все пробелы заменены на знаки абзаца, удалены специфические символы верстки. Целью обработки было создание словника каждой страницы. В результате получилось 8 файлов формата rtf. С помощью программы AndrewTools [Филиппович А.Ю., 2002] были созданы частотные словники каждой страницы и последовательно нескольких страниц (слитые словники). Программа позволяет сохранять словники в виде текстового файла и таблицы Paradox. Далее все расчеты производились вручную. Таблицы частотных словников обрабатывались в Word и осуществлялось их сравнение. Для этого соответствующие слова маркировались цветом, осуществлялась сортировка слов и подсчет. Результаты расчетов и сравнений представлены в таблице 2.10.

Из таблицы видно, что общее количество слов на каждой странице примерно одинаково. Среднее количество слов составляет: $n_{i_cp}=269$ слов. Аналогично среднее количество разных слов $n_{pi_cp}=215$ слов.

Если в процессе корректуры не пользоваться spellером, то количество слов, которые просматривает корректор, будет равно общему количеству слов на странице. В среднем это 269 слов.

Для наглядности представим графическую модель страниц Словаря (рисунок 2.8). Данная модель представляет процесс корректуры с использованием spellера и динамическим пополнением его словаря.

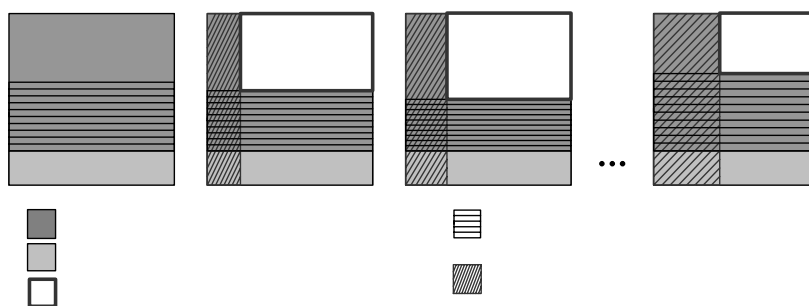


Рисунок 2.8. Графическая модель страниц Словаря.

Первая страница содержит множество слов, часть из которых употребляется несколько раз, такие слова будем называть словоупотреблениями. Количество разных слов в среднем составляет около 80% от общего числа.

Рассматривая разные слова, можно сказать, что около половины этих слов известны Word. В работе корректора эти слова исключаются из рассмотрения, так как они уже входят в состав словаря spellера

На второй странице появляется новая категория слов – слова, которые встречались ранее. Количество этих слов по мере пополнения словаря с каждой последующей страницей растет. Данная группа слов также исключается из рассмотрения, так как эти слова уже входят в словарь spellера. Количество слов, проверяемых корректором, уменьшается с каждой последующей страницей. В процентном соотношении относительно общего количества слов на странице эту тенденцию иллюстрирует экспериментальный график, представленный на рисунке 2.9.

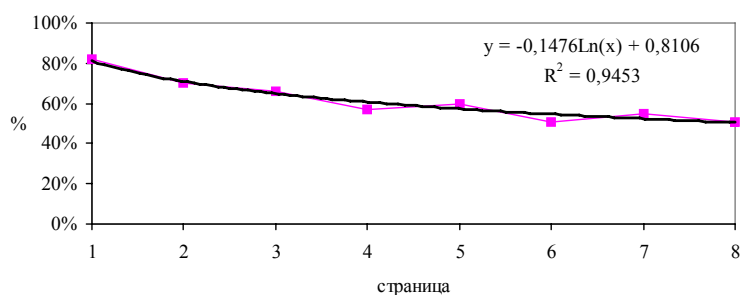


Рисунок 2.9. Соотношение количества слов, проверяемых корректором (в %-ом отношении относительно общего количества слов на странице).

Для формирования общей тенденции распределения частотных характеристик слов формируются аппроксимирующие кривые (при этом вводится допущение, что величины являются непрерывными). При построении кривой используется метод наименьших квадратов, аппроксимация в соответствии с уравнением: $y = c \ln x + b$, где c и b – константы, \ln – функция натурального логарифма. Аналогичным образом строятся аппроксимирующие кривые в последующих исследованиях.

Характеристики последней 570-ой страницы следующие: общее количество слов на странице: 244 (100%) – количество разных слов: 190 (78%); количество разных слов, ранее встречавшихся на этой странице: 142 (59%). В итоге количество слов, которые будет проверять корректор, составляет 190-142=48 слов (20%).

В процессе проведения исследования для каждой страницы было выявлено соотношение слов, известных и не известных Word из числа ранее встречавшихся (рисунок 2.10).

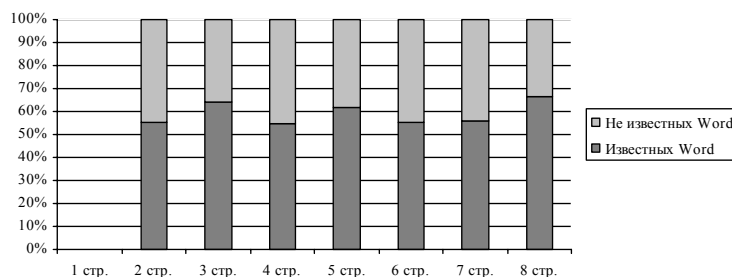


Рисунок 2.10. Соотношение количества ранее встречавшихся слов, известных и не известных Word.

Рассмотрим, какие слова вошли в число неизвестных Word на 8-ой странице из числа тех, что встречались ранее. Большая часть слов – это слова метаязыка – слова, использующиеся для обозначения частей речи, окончания, стилистические пометы и т.п.; а также слова, содержащие буквы, не входящие в современный алфавит, например «Ъ», «Ї» и др.; а также слова, использующие старую форму написания, например, оканчивающиеся на «ть».

Слова, не известные Word, распределены по страницам неравномерно, так, например, если данная страница описывает слова на букву «Ъ», то количество неизвестных слов будет больше, чем на других страницах. Однако, несмотря на колебания соотношений известных и не известных Word слов, из числа ранее встречавшихся, в среднем это соотношение соответствует значению 50/50 для общего количества слов.

Исследование на большой выборке

С целью уточнения количественных характеристик, полученных в результате исследования на малой выборке и характера появления новых слов в тексте CAP, было проведено исследование текста на большой выборке. Предполагается, что в результате данного исследования характер кривой, показывающей количество слов, которое будет проверять корректор, будет аналогичен данным, полученным при исследовании на малой выборке.

Суть исследования аналогична предыдущему. Текст CAP 1-го тома был разбит на 10 частей – выборок по 54 страницы. Каждая последующая выборка сравнивается с предыдущими: вторая с первой, третья с первой и второй и т.д.

Таблица 2.11. Характеристики частотных словариков.

Характеристики сравнения	Выборки				
	1	2	3	4	5
Общее количество слов в словнике	15494	14540	14626	15488	14535
	6	7	8	9	10
	15485	15487	14533	15406	14429
Количество разных слов	1	2	3	4	5
	7275	6642	6758	7068	6208
	6	7	8	9	10
	6872	7029	6523	6906	6489
Количество разных слов без учета регистра	1	2	3	4	5
	6788	6108	6244	6567	5722
	6	7	8	9	10
	6389	6529	6013	6320	5966

Таблица 2.12. Характеристики слитых частотных словариков.

Характеристики сравнения	Выборки				
	1-2	1-3	1-4	1-5	1-6
Количество разных слов	12622	17584	22367	26368	30328
	1-7	1-8	1-9	1-10	
	34509	38141	42057		
Количество разных слов без учета регистра	1-2	1-3	1-4	1-5	1-6
	11585	15989	20244	23760	27240
	1-7	1-8	1-9	1-10	
	30882	33975	37282	40268	

Таблица 2.13. Количество ранее встречавшихся слов.

Количество разных слов ранее встречавшихся	Выборки				
	2	3	4	5	6
С учетом регистра	1295	1797	2287	2207	2913
	7	8	9	10	
	2849	2892	2990	2950	
Без учета регистра	2	3	4	5	6
	1311	1842	2314	2205	2909
	7	8	9	10	
	2889	2920	3013	2980	

Последовательность проведения исследования следующая. Тексты выборок были обработаны в Word с помощью замен: были удалены все знаки препинания, все пробелы были заменены на знаки абзаца, были удалены специфические символы верстки. Целью обработки было создание словарика каждой выборки. В результате получилось 10 файлов формата rtf. Далее были созданы таблицы частотных словариков каждой выборки и последовательно нескольких выборок (слитые словники). Характеристики частотных словариков представлены в таблице 2.11, а слитых словариков – в таблице 2.12. Среднее количество слов каждой выборки 15002.

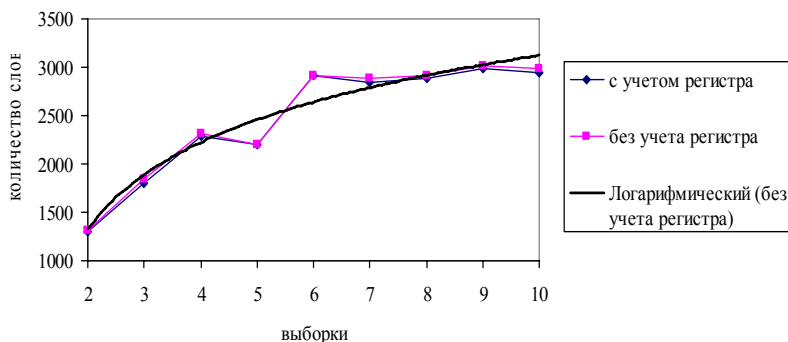


Рисунок 2.11. Рост количества ранее встречавшихся слов в выборках.

С помощью системы запросов в Paradox осуществлялось сравнение таблиц частотных словников. Согласно формальной модели корректуры, представленной параграфе 2.1, необходимо было найти количество слов, ранее встречающихся в предыдущей выборке. Для этого надо найти пересечение множеств этих слов. В исследовании рассматривались словники с учетом регистра и без учета регистра. Из таблицы 2.13 видно, что количество ранее встречавшихся слов в каждой последующей выборке постоянно растет (рисунок 2.11).

Сравним результаты исследования частотных характеристик слов на большой и малой выборках, представив характеристики сравнения в процентном соотношении относительно общего количества слов (таблица 2.14).

В данном случае количество слов в выборке значительно больше, чем в исследовании на малой выборке текста, при этом доля разных слов в выборке значительно меньше и составляет в среднем примерно 42% (для сравнения на одной странице текста 80% разных слов), характер экспериментальной кривой такой же, а количество слов, проверяемых корректором, уменьшается по мере пополнения словаря spellера (рисунок 2.13).

Таблица 2.14. Характеристики сравнения в процентном соотношении (относительно общего количества слов).

Характеристики сравнения	Выборки				
	1	2	3	4	5
Общее количество слов	15494	14540	14626	15488	14535
	6	7	8	9	10
	15485	15487	14533	15406	14429
Количество разных слов	44%	42%	43%	42%	39%
	6	7	8	9	10
	41%	42%	41%	41%	41%
Количество разных слов ранее встречавшихся на странице	1	2	3	4	5
		9%	13%	15%	15%
	6	7	8	9	10
	19%	19%	20%	20%	21%
Количество слов, проверяемых корректором	1	2	3	4	5
	44%	33%	30%	27%	24%
	6	7	8	9	10
	22%	23%	21%	21%	20%

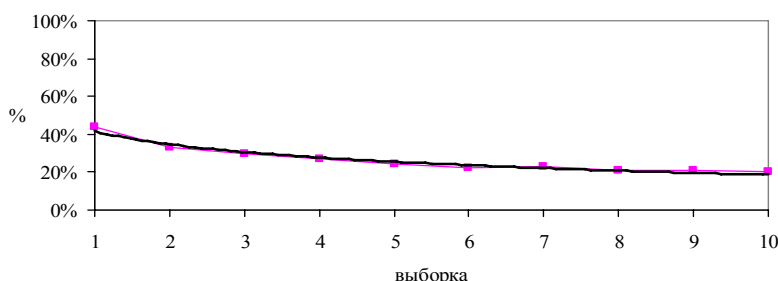


Рисунок 2.12. Соотношение количества слов, проверяемых корректором (относительно общего количества слов).

Таблица 2.15. Характеристики сравнения в процентном соотношении (относительно количества разных слов).

Характеристики сравнения	Выборка, страницы									
	1	2	3	4	5	6	7	8	9	10
<i>На большой выборке</i>										

Количество разных слов, ранее встречавшихся на странице		21	30	35	39	46	44	49	48	50
Количество слов, проверяемых корректором	100	79	70	65	61	54	56	51	52	50
<i>На малой выборке</i>										
Количество разных слов, ранее встречавшихся на странице		12	18	27	26	36	31	35		
Количество слов, проверяемых корректором	100	88	82	73	74	64	69	65		

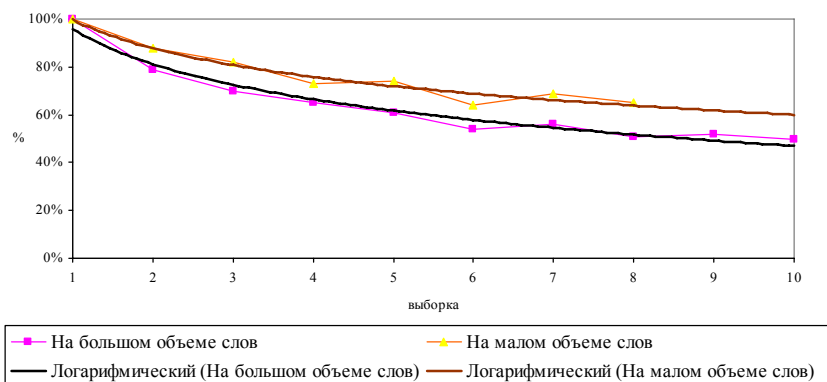


Рисунок 2.13. Соотношения

количества слов, проверяемых корректором (относительно количества разных слов).

Данные исследования на большой выборке подтверждают ранее сделанные, а также могут быть использованы в случае поэтапного ввода текста Словаря.

2.4. Итоги исследований эффективности процедур корректуры

Подведем итоги проведенных исследований методик корректуры с использованием словаря spellera и без него.

Время корректуры текста традиционным методом определяется следующим выражением:

$$T_k^t = \sum_{i=1}^m t_{ki}^t = \sum_{i=1}^m n_i \cdot t_{cp} + \sum_{i=1}^m n_{oi} \cdot t_u,$$

где: m – количество страниц всего текста, t_{ki}^t – время корректуры i -ой страницы текста.

$$t_{ki}^t = n_i \cdot t_{cp} + n_{oi} \cdot t_u,$$

где: t_{cp} – время сравнения слова, t_u – время исправления ошибки, n_i – количество слов на i -ой странице, n_{oi} – количество ошибок на i -ой странице.

Согласно проведенному исследованию, в САР количество слов n_i на каждой странице мало изменяется и составляет в среднем около 269 слов. Считая, что ошибки распределены равномерно, среднее количество ошибок на странице будет равно 13,3 ($\approx 5\%$). Время сравнения слова и исправления в нем ошибки неизвестно. Будем считать, что время исправления ошибки в K раз больше времени сравнения слова, тогда, обозначив время сравнения как t , получим: $t_{cp} = t, t_u = Kt$.

$$t_{ki}^t = n_i \cdot t_{cp} + n_{oi} \cdot t_u = n_i \cdot t + n_{oi} \cdot Kt = n_i t + n_{oi} 0,05 Kt.$$

В итоге для средних значений количества слов и ошибок на странице получим, что

$$T_k^t = (570 \cdot 269)t + (570 \cdot 13,3)Kt = 153330t + 7581Kt$$

Время корректуры текста с использованием словаря spellera определяется следующим выражением:

$$T_k^a = \sum_{i=1}^m t_{ki}^a,$$

$$t_{ki}^a = n_{нов_i} \cdot t_{cp} + n_{oi} \cdot t_u = n_{нов_i} \cdot t + n_{oi} \cdot Kt,$$

где $n_{нов\ i}$ – количество новых (неизвестных) слов на i -ой странице, т.е. количество слов, проверяемых корректором.

В результате исследования для первых восьми страниц был получен экспериментальный график изменения количества новых слов – слов, проверяемых корректором, по мере пополнения словаря spellera:

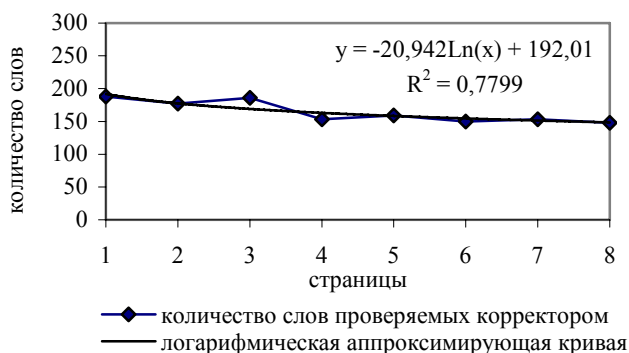


Рисунок 2.14. Количество слов, проверяемых корректором для страниц 1-8.

Для оценки общего количества проверяемых слов при использовании автоматизированной технологии корректуры была построена аппроксимирующая функция. Логарифмическое аппроксимирующее уравнение этой функции имеет вид: $y = -20,94 \cdot \ln x + 192,01$.

Для последующих страниц применим линейную аппроксимацию.



Рисунок 2.15. Количество слов, проверяемых корректором для страниц 8-570.

В этом случае уравнение функции имеет вид: $y = a \cdot x + b$. Известны следующие значения $xy = [8, 148], [570, 48]$. Решая простую систему уравнений, получим:

$$\begin{cases} 8a + b = 148 \\ 570a + b = 48 \end{cases} \text{ получим } a = -\frac{100}{562} \cong -0,178, \quad b = \frac{83976}{562} \cong 149,423.$$

Тогда уравнение прямой: $y = -0,178 \cdot x + 149,423$.

Исходя из этого получим:

а) на промежутке от 1 до 8 $y = -20,94 \cdot \ln x + 192,01$.

б) на промежутке от 9 до 570 $y = -0,18 \cdot x + 149,42$.

Проинтегрируем соответствующие выражения по заданным отрезкам:

$$Y = \int_{x=1}^{x=8} (-20,94 \cdot \ln x + 192,01) dx + \int_{x=9}^{x=570} (-0,18 \cdot x + 149,42) dx$$

$$Y = 1142 + 54934 = 56076$$

Эта величина соответствует количеству новых слов: $\sum_{i=1}^m n_{нов\ i} \cong 56076$

В итоге получим следующее выражение (при условии одинакового среднего времени на исправление ошибок):

$$T_k^a = \sum_{i=1}^m n_{нов\ i} \cdot t + \sum_{i=1}^m n_{oi} \cdot Kt = 56076t + 7581Kt$$

Сравним полученные результаты, вычислив, насколько время автоматизированной корректуры отличается от традиционной, по формуле: $\Delta T_k = 1 - T_k^a / T_k^t$.

При $K=1$ $T_k^t = 160911t$, $T_k^a = 63657t$, $\Delta T_k = 1 - 0,396 = 0,604$;

а при $K=10$ $T_k^t = 229140t$, $T_k^a = 131886t$, $\Delta T_k = 1 - 0,576 = 0,424$

Сравнения позволяют сделать вывод об эффективности методики корректуры с использованием словаря spellera. В случае использования словаря spellera количество слов, сравниваемых корректором, уменьшается и по мере пополнения словаря на последней странице достигает ~20% общего объема.

Эффективность той или иной методики корректуры зависит от соотношения величин времени сравнения слова и времени исправления ошибки. В случае их равенства (K=1) суммарный выигрыш времени корректуры может достигнуть ≈60,4%, а при K=10 он равен ≈42,4%.

Таблица 2.16. Характеристики дополнительных страниц.

Характеристики сравнения	Страницы									
	63	125	188	251	313	377	440	553	570	
Общее количество слов на странице	248	285	254	245	270	250	268	302	244	
Количество разных слов	175	180	197	155	201	187	208	223	190	
Общее количество ранее встречавшихся на странице слов	136	206	167	169	199	188	189	238		
Количество разных слов, ранее встречавшихся на странице	80	109	117	87	135	131	133	165	142	
Количество слов, проверяемых корректором	95	71	80	68	66	56	75	58	48	
<i>В %-ом отношении относительно общего количества слов на странице</i>										
Количество разных слов	71%	63%	78%	63%	74%	75%	78%	74%	78%	
Общее количество ранее встречавшихся на странице слов	55%	72%	66%	69%	74%	75%	71%	79%		
Количество разных слов, ранее встречавшихся на странице	32%	38%	46%	36%	50%	52%	50%	55%	59%	
Количество слов, проверяемых корректором	38%	25%	31%	28%	24%	22%	28%	19%	20%	

При построении кривой изменения количества слов, проверяемых корректором на страницах с 9 по 570, была принята линейная аппроксимация, прямая была построена по характеристикам двух страниц (8-ой и последней).

Для уточнения характера кривой рассмотрим промежуточные значения количества проверяемых слов на периоде с 8 по 570 страницу. При этом данный отрезок был разделен на промежутки по 125 полос (63 страницы). Характеристики значений страниц приведены в таблице 2.16.

Из таблицы видно, что количество слов, проверяемых корректором, изменяется от 38% до 20% относительно общего количества слов на странице. Количество слов, проверяемых корректором исходя из промежуточных значений на промежутке с 9 по 570 страницу, составляет 41160. В случае использования линейной аппроксимации без учета промежуточных значений количество слов составляет 54934. Разница в этих данных составляет 13774 (около 25%).

С учетом этих данных время корректуры будет равно:

$$T_k^a = 42303t + 7581Kt$$

Вычислим насколько время автоматизированной корректуры отличается от традиционной, по формуле:

$$\Delta T_k = 1 - T_k^a / T_k^t$$

При K=1 $T_k^a = 49884t$ суммарный выигрыш времени корректуры может достигнуть ≈69%, а при K=10

$$T_k^a = 118113t \text{ выигрыш времени корректуры } \approx 48,5\%.$$

Для оценки общего количества слов на промежутке с 9 по 570 страницу построим аппроксимирующую кривую с учетом промежуточных значений. В качестве метода аппроксимации используем метод наименьших квадратов и линейную зависимость (рисунок 2.16).

Уравнение аппроксимирующей прямой имеет вид: $y = -0,11x + 10,84$.

Тогда:

$$Y = \int_{x=1}^{x=8} (-20,94 \cdot \ln x + 192,01) dx + \int_{x=9}^{x=570} (-0,11 \cdot x + 10,84) dx = 44015$$

Эта величина соответствует количеству новых слов: $\sum_{i=1}^m n_{\text{нов } i} \cong 44015$

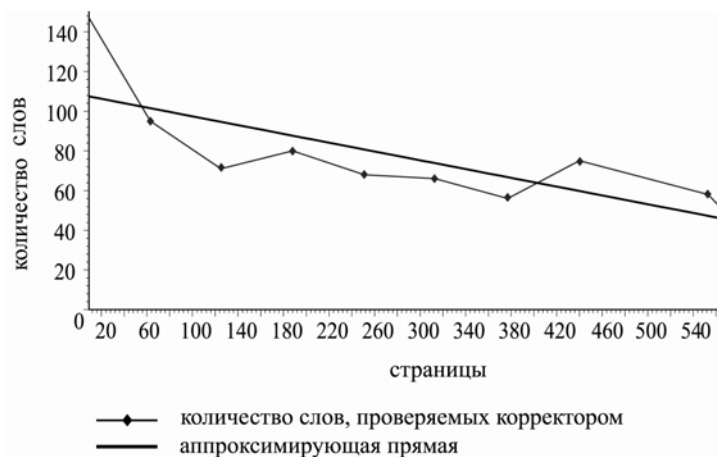


Рисунок 2.16. Соотношение количества слов, проверяемых корректором для страниц 8-570 с учетом промежуточных значений.

В итоге получим следующее выражение, при условии одинакового среднего времени на исправление ошибок:

$$T_k^a = \sum_{i=1}^m n_{нов\ i} \cdot t + \sum_{i=1}^m n_{о\ i} \cdot Kt = 44015t + 7581Kt$$

При $K=1$ $T_k^a = 51596t$, суммарный выигрыш времени корректуры может достигнуть $\approx 68\%$, а при $K=10$ $T_k^a = 119825t$ и выигрыш времени корректуры $\approx 47,7\%$.

Таким образом с учетом уточнения экспериментальной кривой на промежутке с 8 по 570 страницу объем слов, проверяемых корректором уменьшился на 25%. Это дает выигрыш времени еще на $\approx 6-9\%$.

Оценивая полученные показатели, следует отметить ряд допущений, которые были приняты в формальной модели корректуры. Во-первых, было принято, что ошибки распределены по тексту равномерно, поэтому количество ошибок на каждой странице постоянно. Во-вторых, рассматривались только орфографические ошибки, не рассматривались ошибки пунктуации и связанные с нарушением правил верстки. В данную модель не входят также ошибки в словах, входящих в состав словаря spellera.

Полученные результаты, однако, позволяют рекомендовать методику корректуры с использованием словаря spellera при первой читке. Для обнаружения всех остальных ошибок целесообразно сохранить традиционную методику корректуры (при второй и третьей читке).