Корректура текста

Технологии корректуры

Лекция №9

Лингвистическое обеспечение АСОИУ

К.т.н., доцент Филиппович Анна Юрьевна

Корректура

- ***Корректура*** – совокупность процессов, назначением которых является исправление ошибок и нарушений технических правил в наборе.

Факторы, влияющие на процесс корректуры

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

Корректура

- Сегодня для подготовки текстов используются различные программы верстки и текстовые редакторы.
- А в качестве средств автоматизации корректурных процессов выступают различные встроенные функции проверки текста на наличие орфографических, синтаксических и стилистических ошибок.
- Одна из таких функций – функция ***спеллер*** (speller – сокращение от spelling checker – программа поиска опечаток, корректор).

Инструментарий корректора

- Печатные словари.
- Электронные лексикографические ресурсы, в числе которых:
  - локальные электронные словари;
  - интернет-порталы;
  - словарные базы данных;
  - встроенные в текстовые редакторы и издательские системы орфо- и грамматические редакторы;
  - программы спеллеры.

Особенность современных программ проверки текстов

- Особенность современных программ проверки текстов является их **ориентация на современную общеупотребительную лексику**, что затрудняет их использование для специфических, старинных текстов.

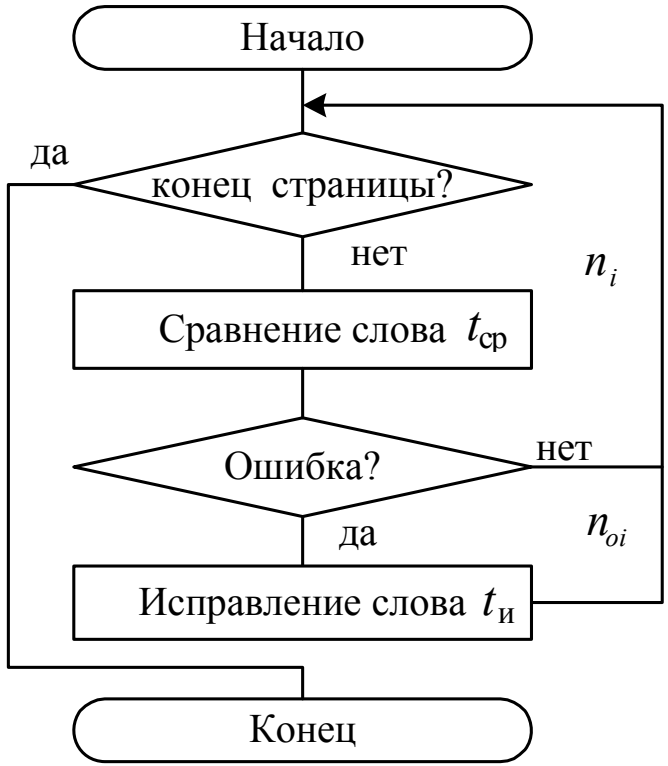
Традиционная технология корректуры

$$t_{к\,i} = n_i \cdot t_{ср} + n_{o\,i} \cdot t_u$$

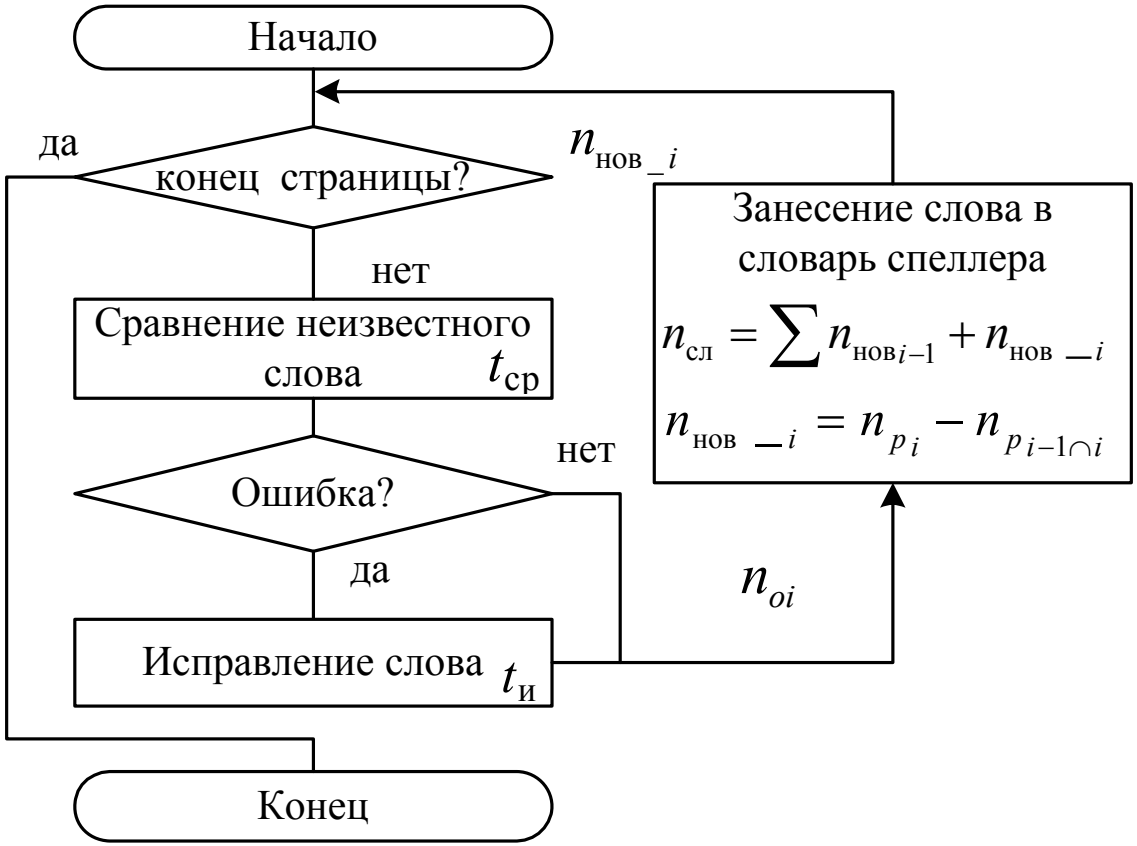
где: *tср* – время сравнения слова,
tи – время исправления ошибки;
ni – общее количество слов на *i*-ой странице,
noi – количество ошибок на *i*-ой странице.

$$T_k^t = \sum_{i=1}^{m} t_{к\,i} = \sum_{i=1}^{m} n_i \cdot t_{ср} + \sum_{i=1}^{m} n_{o\,i} \cdot t_u$$

где *m* – количество страниц всего текста.



Автоматизированная технология корректуры



Автоматизированная технология корректуры

$$t_{к\,i} = n_{нов\_i} \cdot t_{ср} + n_{о\,i} \cdot t_{и}$$

где *пнов_i* – количество новых слов на i-ой странице,
noi – количество ошибок на i-ой странице.

Количество новых слов – занесенных в словарь:

$$n_{сл} = \sum n_{нові-1} + n_{нові}$$

$$n_{нов\,i} = n_{p\,i} - n_{p\,i-1 \cap i}$$

где *npi* – количество разных слов на i-ой странице (неповторяющихся на странице),

$n_{p\,i-1 \cap i}$ - количество общих разных слов i-ой и предыдущей (i-1) странице.

Эффективность технологии корректуры

- Эффективность той или иной технологии корректуры будем определять исходя их времени, затрачиваемом на корректуру текста.
- В формальной модели корректуры фигурируют два вида параметров: **время**, затрачиваемое, на ту или иную деятельность и **количественные характеристики**.
- Проанализируем количественные характеристики на примере корректуры Словаря Академии Российской 1789-1794 гг.

Исследование количества ошибок

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

Характеристики сравнения (кол-во)	**Введенный текст**	**Вычитанный текст**
Всего записей	6092	6103
Всего неповторяющихся записей	6078	6094
Всего неповторяющихся слов	6031	6049
Одинаковых записей	5499	
Одинаковых неповторяющихся записей	5477	
Одинаковых неповторяющихся слов	5571	
Ошибок в неповторяющихся записях	601	
Ошибок в неповторяющихся словах	460	
Ошибок в номерах колонок	108	
Отсутствующих записей	11	
Отсутствующих номеров колонок	33	

Исследование количества ошибок

- 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

Ошибки, связанные со старинной лексикой и грамматикой

Описание ошибки	Примеры		Кол-во ошибок
	Ошибки	Исправления	
Отсутствие Ъ на конце	Абшит Бекет Вдовец	АбшитЪ БекетЪ ВдовецЪ	17
ѣе→ѣ	Бѣлоручка Бѣшуся НабѣегЪ	Бѣлоручка Бѣшуся НабѣгЪ	20
ий→ї, ий→їй ие→їе	Повязывание Вороniий Провѣщаниїе Бальсамический Воинский Орудие Збывание	Повязыванїе Воронїй Провѣщанїе Бальсамическїй Воинскїй Орудїе Збыванїе	16
Старинное написание слов	БадьянЪ Сибирский Вельможество Испровергаютпся Оружеборимый Подбираюсь Ублаженїе	БадьянЪ Сибирской Вельможпво Испровергаюся Оружеборный Подбиваюся Ублажанїе	≈20

Анализ систематических ошибок

Ошибки, обусловленные особенностями графем шрифта

Символы	Примеры		Кол-во ошибок
	Ошибки	Исправления	
п←→ш	АскишЪ Баронсшво Волишель ВоропникЪ Наблопняю Обеппалость	АскипЪ Баронспво Волипель ВоропникЪ Наблопняю Обеппалость	13
ш←→щ	Блудяшїе огни Вѣщанїе Всевысочайще	Блудяшїе огни Вѣшанїе Всевысочайше	3
ѣ←→Ъ	Аптекаревѣ Барвенокѣ Безѣизбѣжно ВнЪ ВывЪвки ЕдиновЪрїе	АптекаревЪ БарвенокЪ БезЪизбѣжно Внѣ Вывѣвки Единовѣрїе	32
ь←→Ъ	Билїардь ВѣстЪ НеворотЪ	БилїардЪ Вѣсть Неворотъ	5
л←→д	АспиловЪ Водохранидище	АспидовЪ Водохранилище	3

Анализ систематических ошибок

Другие систематические ошибки (технические)

Описание ошибки	Примеры		Кол-во ошибок
	Ошибки	Исправления	
ю→.	АзЪ мѣстоию Балакирью Изневѣстью	АзЪ мѣстои. Балакирь. Изневѣсть.	12
Прописные буквы после точки .␣Н←→.␣н	Аа. Межд АльтЪ. Скрыпка ВьюрокЪ. Пташка	Аа. Межд АльтЪ. Скрыпка ВьюрокЪ. Пташка	35

Анализ систематических ошибок

Другие систематические ошибки

Описание ошибки	Примеры		Кол-во ошибок
	Ошибки	Исправления	
␣Ї, ␣Їе, ␣ЇЙ, ␣Ї␣е, ␣Ї␣Й	Б їйца Баснослов їе Бомбардирск їй Благосовѣт ї е Боярск ї й	Бїйца Баснословїе Бомбардирскїй Благосовѣтїе Боярскїй	49
і→ї	Визжаніе Забѣганіе Завоеваніе	Визжанїе Забѣганїе Завоеванїе	19
Повторение символов	АрканЪЪ Брезззгунька БѣБѣлена	АрканЪ Брезгунька БѣБѣлена	25
Перестановка символов	Бабки вольчи Сповѣлеваю	Бабки волчьи Сповелѣваю	2

Исследование частотных характеристик слов

Графическая модель страниц Словаря

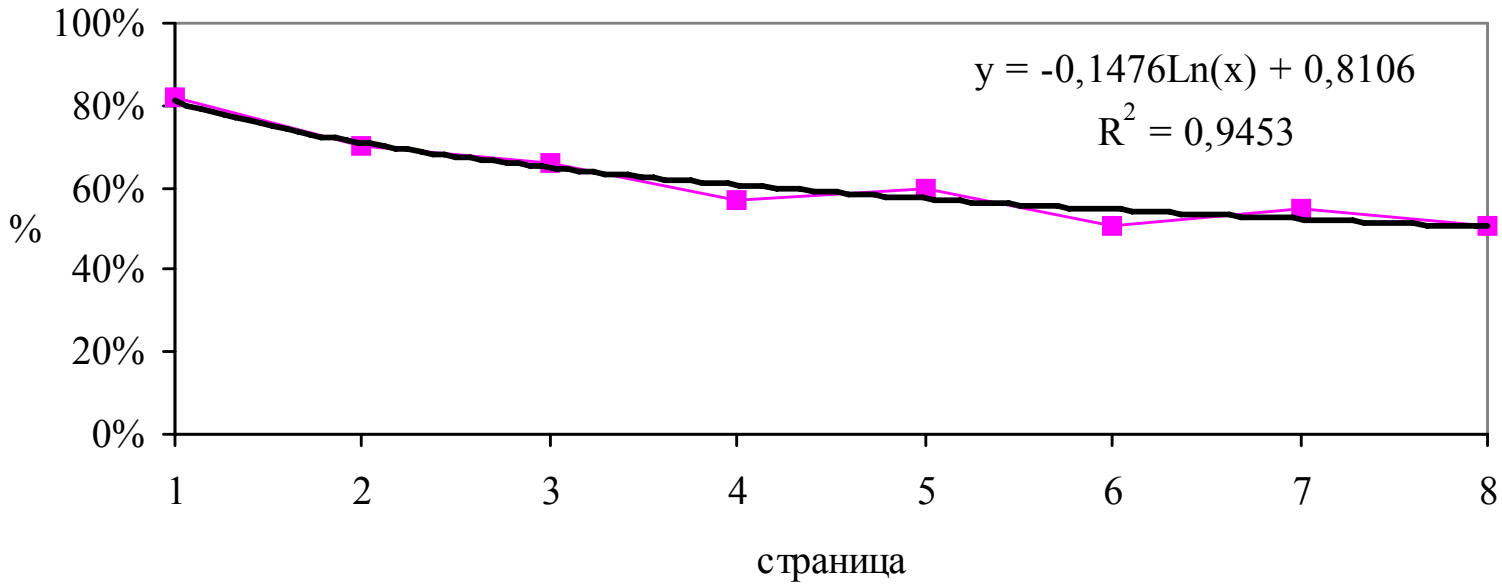
Исследование частотных характеристик слов

Характеристики страниц 1-8

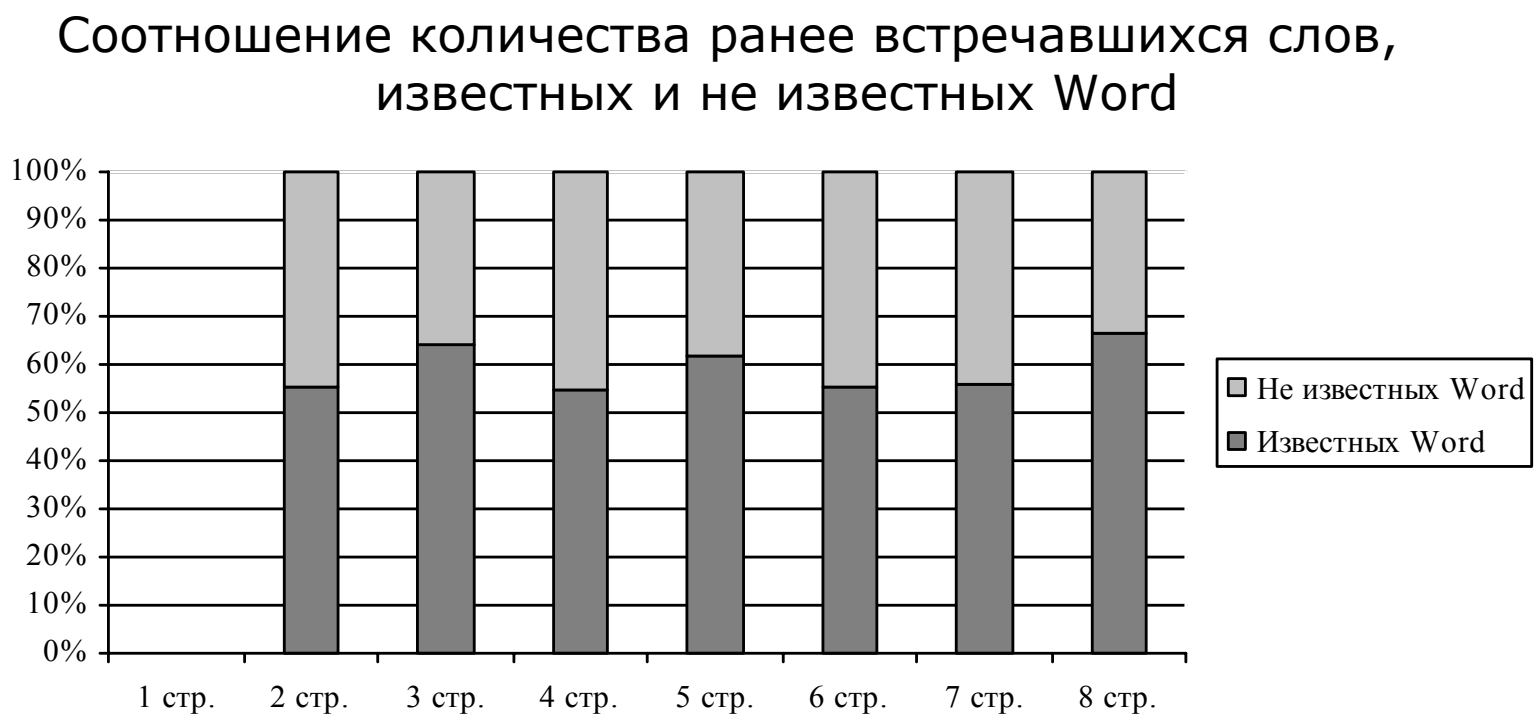
Характеристики сравнения	**Страницы**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Общее количество слов на странице	228	256	279	268	265	294	276	288
Количество разных слов	188	201	227	211	215	233	222	226
Общее количество ранее встречавшихся на странице слов		51	86	103	101	130	107	125
Количество разных слов ранее встречавшихся на странице		24	41	58	56	83	69	78
Количество слов проверяемых корректором	188	177	186	153	159	150	153	148

Исследование частотных характеристик слов

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

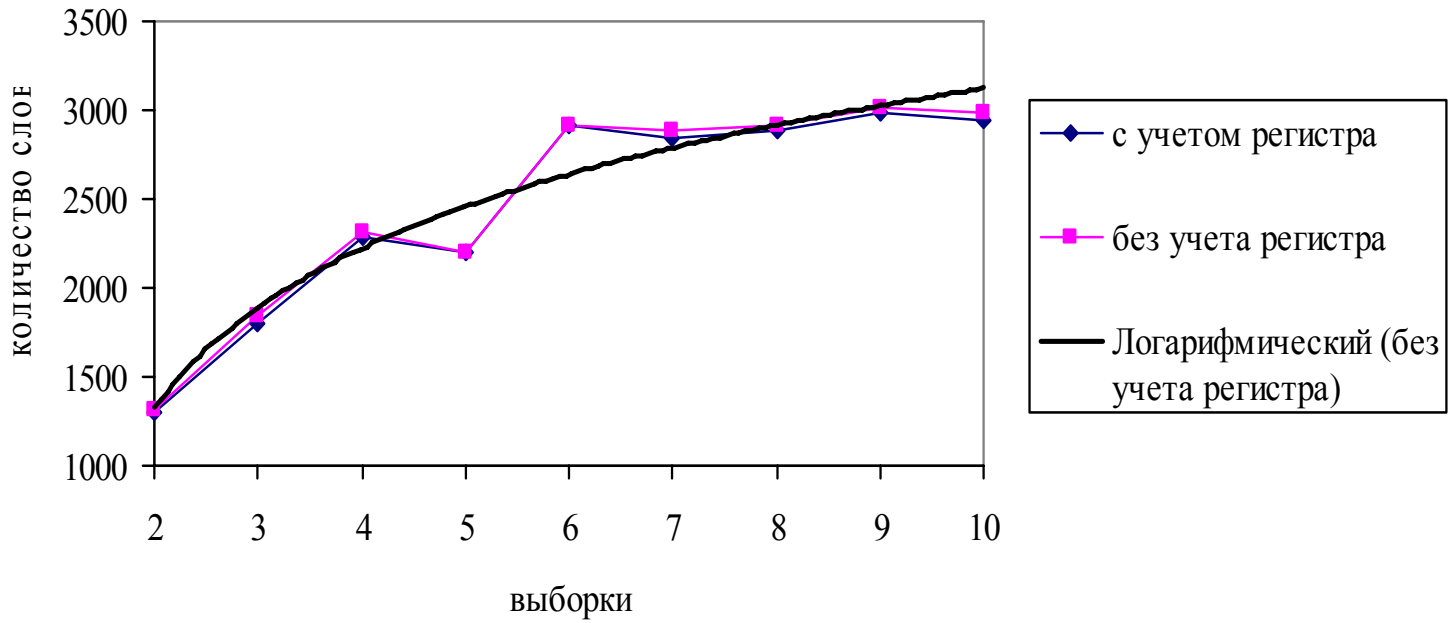


Исследование частотных характеристик слов



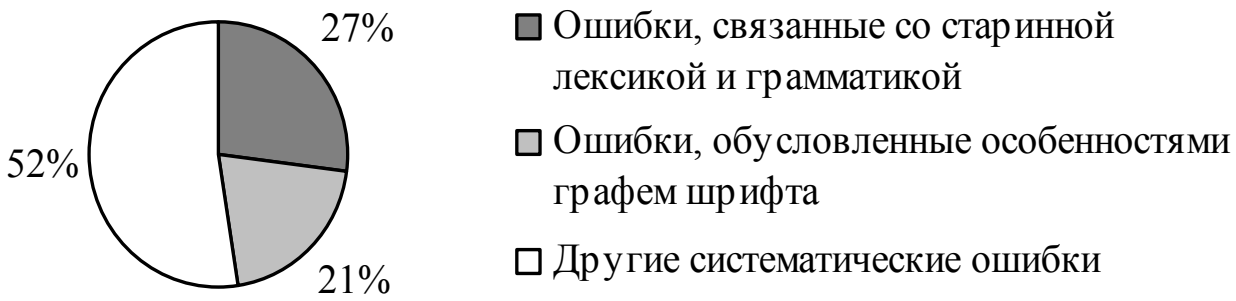
Исследование частотных характеристик слов

Рост количества ранее встречавшихся слов в выборках



Эффективность технологии корректуры

- Для оценки общего количества проверяемых слов при использовании автоматизированной технологии корректуры для 1-8 страниц была построена аппроксимирующая функция:



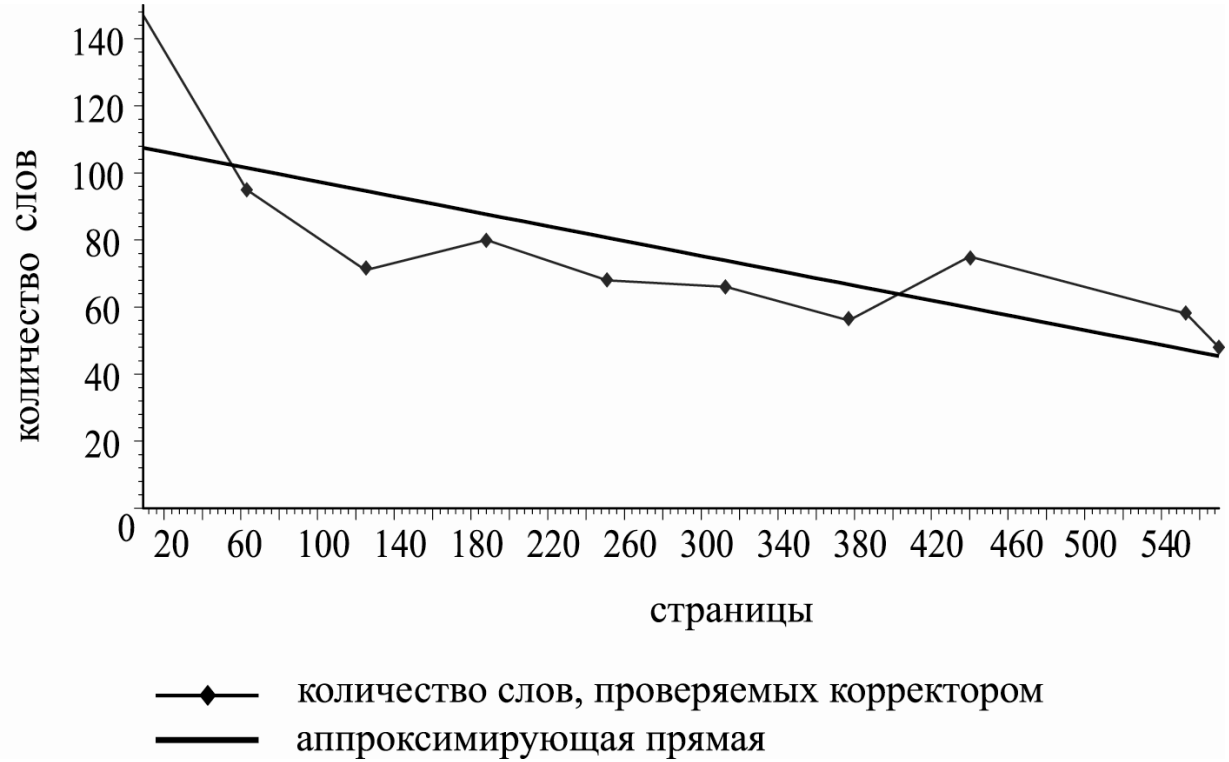
Эффективность технологии корректуры

- 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.

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

Эффективность технологии корректуры

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



Эффективность технологии корректуры

- Количество слов, проверяемых корректором при использовании автоматизированной технологии корректуры определяется следующим выражением:

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

Эта величина соответствует количеству новых слов:

$$\sum_{i=1}^{m} n_{нов\ i} \cong 44015$$

Эффективность технологии корректуры

- Будем считать, что время исправления ошибки в K раз больше времени сравнения слова, тогда, обозначив время сравнения как *t*, получим:
  tср = t, tи = Kt.

Эффективность технологии корректуры

- Традиционная технологий корректуры:

$$T_k^t = \sum_{i=1}^{m} t_{k\,i}^t = \sum_{i=1}^{m} n_i \cdot t + \sum_{i=1}^{m} n_{o\,i} \cdot Kt$$

$$T_k^t = 153330t + 7581Kt$$

- Автоматизированная технологий корректуры:

$$T_k^a = \sum_{i=1}^{m} t_{k\,i}^a = \sum_{i=1}^{m} n_{нов\ i} \cdot t + \sum_{i=1}^{m} n_{o\,i} \cdot Kt$$

$$T_k^a = 44015t + 7581Kt$$

Эффективность технологии корректуры

- Сравнение технологий корректуры:

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

При К=1 , суммарный выигрыш времени корректуры может достигнуть ≈68%,

а при К=10 и выигрыш времени корректуры ≈47,7%.

Эффективность технологии корректуры

- 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 спеллера.