Московский государственный университет печати имени Ивана Фёдорова

Кафедра медиасистем и технологий

Анна Юрьевна Филиппович
Юрий Николаевич Филиппович

Технологии корректуры текста

Лекции по дисциплине
"ИНТЕГРИРОВАННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ В МЕДИАИНДУСТРИИ"

Для студентов,
осваивающих образовательные программы
подготовки бакалавров и магистров по направлениям
230400 «Информационные технологии и системы»,
230100 «Информатика и вычислительная техника»

Москва, 2013

Корректура

Основные понятия

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

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. 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.

Корректурный знак	Назначение знака	Применение в тексте
Z	Начать текст с абзаца	Z *Просим выделить необходимые средства для ремонта.*
(знак начала и конца переноса)	Набрать в подбор с текстом (без абзаца)	*Инженер Петров освобожден от занимаемой должности по собственному желанию. Основанием послужило личное заявление г-на Петрова.*
L Г T ⊥ H Н L L	Вписать пропущенные буквы	*Необходимо восстановить данный кумент.* Г *до*
V Λ V̄ Λ̄ V	Вписать недостающие или ошибочно напечатанные слова	*Ошибочно напечатанные буквы, слоги, слова перечеркивают и ними надписывают нужные.* V̄ *над*

Рис.1. Примеры корректурных знаков.

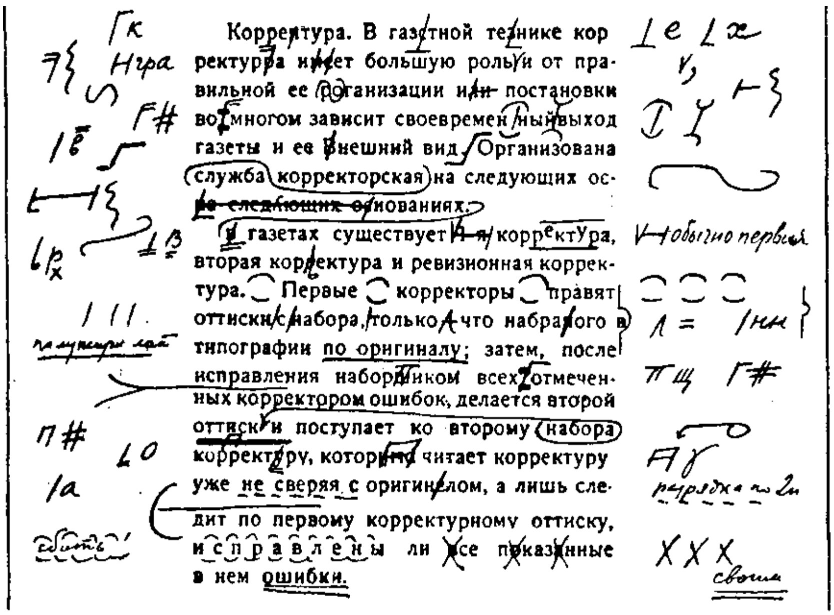
Коррертура. В газатной теенике кор ректурра ииеет большую рольи от правильной ее роганизации и и постановки вомногом зависит своевремен ный выход газеты и ее внешний вид Организована служба корректорская на следующих ос-
но следующих основаниях.
В газетах существует и я корректура, вторая корректура и ревизионная корректура. Первые корректоры правят оттиски с набора, только что набраного в типографии по оригиналу; затем, после исправления наборщиком всех отмеченных корректором ошибок, делается второй оттиски поступает ко второму набора корректуру, которые читает корректуру уже не сверяя с оригиналом, а лишь следит по первому корректурному оттиску, исправлены ли все показанные в нем ошибки.

Рис. 2. Фрагмент корректурной правки.

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.

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.).

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.

Словари

Рассмотрим словари, которые могут использовать корректоры. Существуют два основных типа:

Лингвистические словари. Объектом описания лингвистических словарей являются языковые единицы: слова, устойчивые словосочетания, словоформы, морфемы и др. Пример словарной статьи из **лингвистического словаря**:

СУРОК, -р к а, м. Небольшой грызун сем. бельчьих, живущий в норах и зимой впадающий в спячку.

Энциклопедические словари – научные или научно-популярные справочные издания, представляющие собой систематизированный свод знаний в каких-либо областях. Объектом их описания служат научные (реже обиходные) понятия, термины, исторические события, персоналии, географические реалии и т.д.

Пример словарной статьи из **энциклопедического словаря**:

СУРКИ, род млекопитающих сем. бельчьих. Длина тела до 60 см, хвоста менее 1/2 длины тела. 13 видов, в Сев. полушарии (исключая пустыни и тундры); в России неск. видов. Объект промысла (мех, жир, мясо). Могут быть носителями возбудителя чумы. Нек-рые виды редки, охраняются.

Для корректуры применяются грамматические словари – это словари, которые содержат сведения о морфологических и синтаксических свойствах слова.

Расположение − в прямом или обратном алфавитном порядке. Принципы отбора и объем сведений о слове различны в зависимости от назначения и адресата каждого грамматического словаря.

Морфемные и словообразовательные словари также используются корректором. Словари, показывающие членение слов на составляющие их морфемы, словообразовательную структуру слова, а также совокупность слов с данной морфемой – корневой или аффиксальной. Слова в словообразовательных словарях приводятся с

расчленением на морфемы и с ударением. Морфема (от греч. morphe – форма) – минимальная значимая часть слова.

Орфографические словари, содержащие алфавитный перечень слов в их нормативном написании.

Среди электронных словарей следует выделить проект «РУССКИЕ СЛОВАРИ». Он предназначен для всех, кто интересуется русским языком - как родным или как иностранным, для учащихся средней и высшей школы, а также для специалистов, профессионально занимающихся лингвистикой или преподаванием русского языка. Он содержит общедоступную лингвистическую информацию разного типа. Словарная база сайта содержит 21 том основных интерактивных лингвистических словарей, многие из которых входят в золотой фонд отечественной лексикографии. Режим доступа: http://www.slovari.ru/

А также другие:

- Яндекс-словари содержат 11 словарей русского языка, 86 энциклопедий и переводной словарь (7 языков). http://slovari.yandex.ru/
- Словари и энциклопедии на Академике http://dic.academic.ru/
- «Кругосвет» – универсальная энциклопедия http://www.krugosvet.ru/
- Википедия – свободная энциклопедия http://ru.wikipedia.org/
- Словари на «Рубриконе» – река информации http://www.rubricon.com/
- «Мир словарей» – коллекция словарей и энциклопедий http://mirslovarei.com/
- Мир энциклопедий http://www.encyclopedia.ru/index.html

Электронные переводные словари:

- Система электронных словарей Lingvo http://www.lingvo.ru/
- Онлайновый переводчик компании ПРОМТ http://www.translate.ru/Rus/
- Электронные словари компании Мультилекс http://www.multilex.ru/
- Переводной словарь Google http://www.google.ru/dictionary?hl=ru

Сегодня для подготовки текстов используются различные программы верстки и текстовые редакторы.

Технологии корректуры

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).

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

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. Оценим эффективность этих методик путем исследования их формальных моделей.

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_{ki}. Оно будет определяться через следующее выражение:

$$t_{кi} = n_i \cdot t_{ср} + n_{oi} \cdot t_u ,$$

где: $t_{ср}$ – 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:

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

где m – количество страниц всего текста.

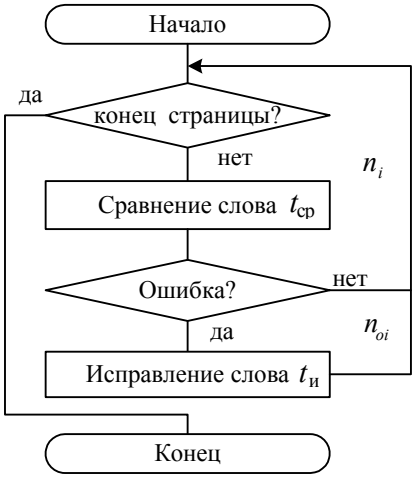

Рис. 3. Традиционная методика корректуры страницы текста.

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.

Предположим, что словарь спеллера 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:

$t_{кi} = n_{нов\_i} \cdot t_{ср} + n_{oi} \cdot t_u$, где $n_{нов\_i}$ – количество новых слов на i-ой странице, n_{oi} – количество ошибок на i-ой странице.

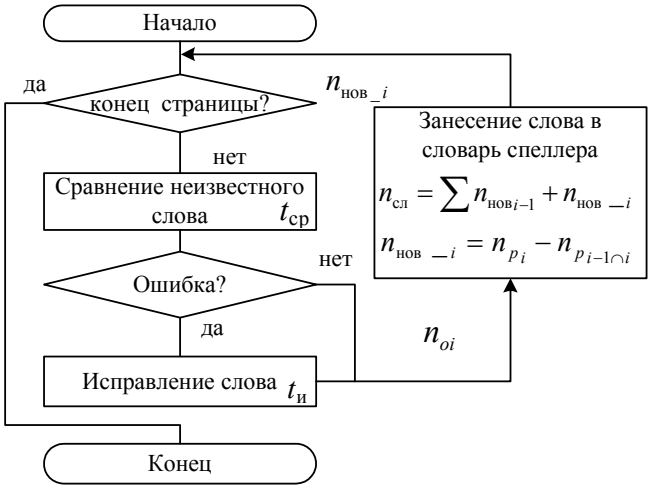

Рис. 4. Автоматизированная методика корректуры страницы текста.

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: <W_Z, K>, где W_Z – слово (словосочетание) Показания, K – номер колонки. Сравнивая таблицы Pok1tOsh и Pok1t мы получили таблицу ошибок.

Таблица 1. 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 слов	5571	
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. 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	Бѣелоручка Бѣешуся НабѣегЪ	Бѣлоручка Бѣшуся НабѣгЪ	**20**
иї→ї, ий→їй ие→їе	Повязывание Вороніїй Провѣщаніе Бальсамический Воинский Орудие Збывание	Повязываніе Воронїй Провѣщанїе Бальсамическїй Воинскїй Орудїе Збываніе	**16**
Старинное написание слов	БадьянЪ Сибирский Вельможество Испровергаются Оружеборимый Подбираюсь Ублаженїе	БадьянЪ Сибирской Вельможство Испровергаюся Оружеборный Подбиваюся Ублажанїе	**≈20**

Часть ошибок связана со старинной славянской лексикой и грамматикой (27%), используемой в Словаре (таблица 2). Это касается и специфических символов – букв старого алфавита, и целых слов, которые уже вышли из употребления. Ярким примером подобной ошибки является отсутствие «Ъ» 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), Испровергаютпся (Испровергаюся), Ублажанїе (Ублаженїе).

Таблица 3. Ошибки, обусловленные особенностями графем шрифта.

Символы	Примеры		Кол-во ошибок
	Ошибки	Исправления	
т←→ш	АскишЪ Баронсшво Волишель ВоротникЪ Наблотняю Обеттталость	АскитЪ Баронство Волитель ВорошникЪ Наблошняю Обетшалость	**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**

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 «т», «ш» и «щ». Клавиши с буквами «ш» и «щ» находятся рядом, поэтому данная ошибка может быть вызвана этим. Примером похожих букв с округлыми элементами являются «ь», «Ъ» и «ѣ».

Согласно исследованию, наибольшее количество ошибок было связано с путаницей «ъ» и «ѣ» (таблица 3), 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»: ЛирЪ – пирЪ; «е» и «с»: Верета – Верста; «л» и «я»: Выл – Выя; «ь» и «б»: Оьладатель – Обладатель.

Другие систематические ошибки составляют около 51%. Среди них технические ошибки набора (таблица 4). 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. Пташка.

Таблица 4. Технические ошибки.

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**

В тексте Показания обнаружены систематические ошибки, которые трудно отнести к какой-либо группе и определить их причины (таблица 5). Самое большое количество ошибок (49) это наличие пробелов в словах, особенно перед символом «ї». Другой пример – это ошибки-опечатки: повторение символов и сочетаний а также их перестановка.

Таблица 5. 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**

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.

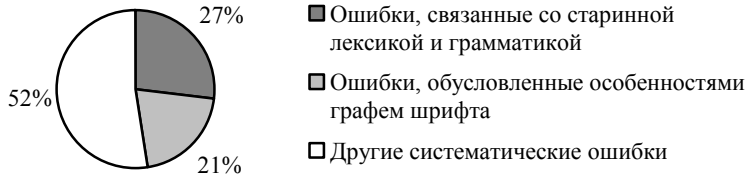

Рис. 5. Соотношение разных видов ошибок.

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→їй, ие→їе, юпся→юся**. Для технических и других систематических ошибок: **і→ї, иї→ї, ъе→ѣ, ␣ї→ї.**

Исследование частотных характеристик слов

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 спеллера).

Таблица 6. Характеристики страниц 1-8.

Характеристики сравнения	**Страница**							
	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		51	86	103	101	130	107	125

встречавшихся на странице слов								
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 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 спеллером, 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 (рис. 6). Данная модель представляет процесс корректуры с использованием спеллера и динамическим пополнением его словаря.

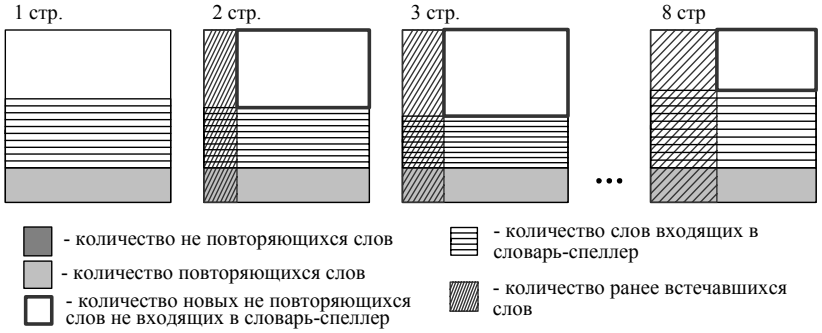

Рис. 6. Графическая модель страниц Словаря.

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 спеллера

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 спеллер. 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 7.

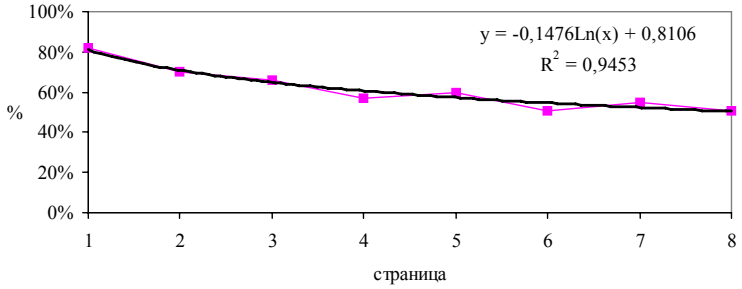

Рис. 7. Соотношение количества слов, проверяемых корректором (в %-ом отношении относительно общего количества слов на странице).

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: $y = c\ lnx + b$, 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 (рис. 8).

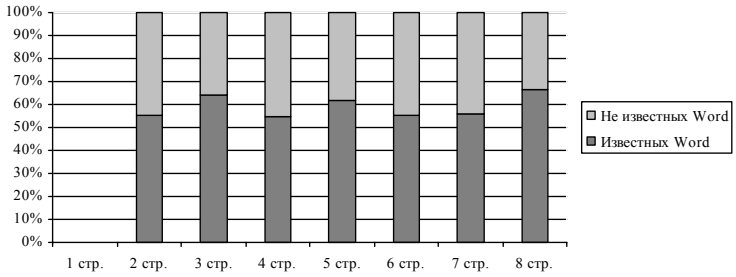

Рис. 8. Соотношение количества ранее встречавшихся слов, известных и не известных Word.

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 «ѣ», «ї» и др.; а также слова, использующие старую форму написания, например, оканчивающиеся на «ъ».

Слова, не известные Word, распределены по страницам неравномерно, так, например, если данная страница описывает слова на букву «ѣ», 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 САР, 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 САР 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 7. Характеристики частотных словников.

Характеристики сравнения	**Выборка**				
Общее количество слов в словнике	**1**	**2**	**3**	**4**	**5**
	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 8. Характеристики слитых частотных словников.

Характеристики сравнения	**Выборка**				
Количество разных слов	**1-2**	**1-3**	**1-4**	**1-5**	**1-6**
	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 9. Количество ранее встречавшихся слов.

Количество разных слов ранее встречавшихся	**Выборка**				
С учетом регистра	**2**	**3**	**4**	**5**	**6**
	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 словника каждой выборки. В результате получилось 10 файлов формата rtf. Далее были созданы таблицы частотных словников каждой выборки и последовательно нескольких выборок (слитые словники). Характеристики частотных словников представлены в таблице 7, а слитых словников – в таблице 8. Среднее количество слов каждой выборки 15002.

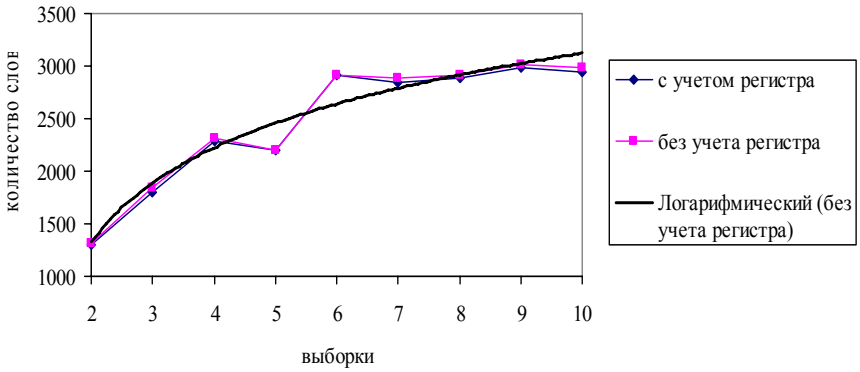

Рис. 8. Рост количества ранее встречавшихся слов в выборках.

С помощью системы запросов в Paradox осуществлялось сравнение таблиц частотных словников. Согласно формальной модели корректуры, 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 9 видно, что количество ранее встречавшихся слов в каждой последующей выборке постоянно растет (рис. 8).

Сравним результаты исследования частотных характеристик слов на большой и малой выборках, представив характеристики сравнения в процентном соотношении относительно общего количества слов (таблица 10).

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 спеллера (рис. 9).

Таблица 10. Характеристики сравнения в процентном соотношении (относительно общего количества слов).

Характеристики сравнения	**Выборка**				
Общее количество слов	**1**	**2**	**3**	**4**	**5**
	15494	14540	14626	15488	14535
	6	**7**	**8**	**9**	**10**
	15485	15487	14533	15406	14429
Количество разных слов	**1**	**2**	**3**	**4**	**5**
	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%

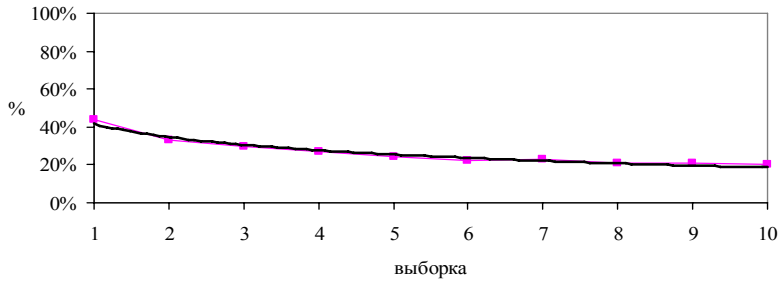

Рис. 9. Соотношение количества слов, проверяемых корректором (относительно общего количества слов).

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

Итоги исследований эффективности процедур корректуры

Подведем итоги проведенных исследований методик корректуры с использованием словаря спеллера 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 (≈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: $t_{cp} = t$, $t_u = Kt$.

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

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 спеллера 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 спеллера:

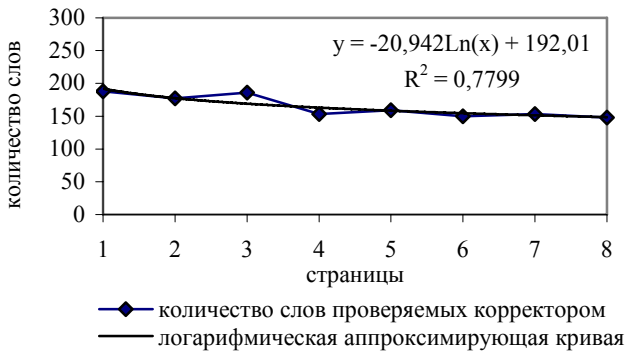

Рис. 10. Количество слов, проверяемых корректором для страниц 1-8.

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.

Рис. 11. 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:

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

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):

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

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

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

а при К=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 спеллера. В случае использования словаря спеллера 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 (К=1) суммарный выигрыш времени корректуры может достигнуть ≈60,4%, а при К=10 он равен ≈42,4%.

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 11.

Таблица 11. 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
В %-ом отношении относительно общего количества слов на странице									
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%

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 К=1 $T_k^a = 49884t$ суммарный выигрыш времени корректуры может достигнуть ≈69%, а при К=10 $T_k^a = 118113t$ выигрыш времени корректуры ≈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 (рис. 12).

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

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

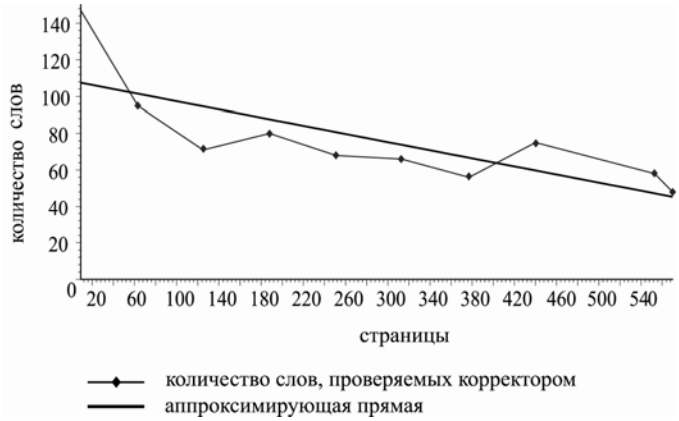

Рис.12. Соотношение количества слов, проверяемых корректором для страниц 8-570 с учетом промежуточных значений.

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

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

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

Т.о. 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 ≈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 спеллера.

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

Литература

Борковский А. Б. Англо-русский словарь по программированию и информатике (с толкованиями) – М.: Рус. яз., 1989. – 335 с.

Волкова Л.А., Решетникова Е.Р. Технология обработки текстовой информации. Часть I. Основы технологии издательских и наборных процессов. Издание второе, исправленное и дополненное: Учебное пособие. М.: Изд-во МГУП, 2002. 306 с.

Гунько С.Н. Демков В.И. Словарь по полиграфии и полиграфической технологии. Понятия и определения. – Мн.: ООО «Космополис-Универсал», 1995. – 230 с.

Методы и системы автоматизированного обнаружения и коррекции текстовых ошибок/ О. Б. Бабко-Малая, В. А. Шемраков. - Л. : БАН, 1987. - 48 с.

Рисс О.В. Что нужно знать о корректуре Маленькое пособие. – М. Книга 1980.

Рыжова Л.А. Корректура : учеб. пособие для студентов учреждений сред. проф. образования, обучающихся по специальности 0206 "Изд. дело". – М. МИПК им. И. Федорова 2005.

Справочное пособие для редакторов и корректоров / [Сост. В. Ю. Лернер, Н. А. Теплякова] . – М. : Медицина, 1984.

Филиппович А.Ю. Исследование эффективности автоматизации корректурных процессов с помощью словаря спеллера при подготовке переиздания Словаря Академии Российской 1789–1794 гг. [Текст] // Проблемы полиграфии и издательского дела. № 4. – М.: Изд-во МГУП, 2007. – С. 102-112.

Филиппович А.Ю. Практические занятия по курсам «Компьютерная лингвистика» и «Семиотика информационных технологий». Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. Выпуск 6, 2005 г.

Филиппович А.Ю. Словарь Академии Российской (1789–1794): информационная технология переиздания. Вступительная статья М.И.Чернышевой. — М.: МГУП, 2008. – 304 с.

Филиппович. А.Ю. Лингвистический редактор Andrew Tools 2000. // Проблемы прикладной лингвистики 2001. Сборник статей / Отв. Ред. А.И. Новиков. – М. «Азбуковник», 2001. – 360 с.