

**Московский государственный технический университет им. Н.Э. Баумана
кафедра "Системы обработки информации и управления"**

**ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ИССЛЕДОВАНИЯ ЯЗЫКА
ПЕЧАТНЫХ ИСТОЧНИКОВ XVIII – НАЧ. XIX ВВ.**

Расчетно-пояснительная записка
курсовой работы по дисциплине
"Семиотика информационных технологий"
студент группы ИУ5-92
Прохоров Артём Вячеславович.

Шифр: 92_15

Преподаватель: к.т.н., доц. А.Ю.Филиппович

Москва, 2010 г.

Содержание.

Введение.....	3
Основная часть.....	4
Задача 1.....	4
Задача 2.....	5
Задача 3.....	8
Задача 4.....	12
Задача 5.....	14
Задача 6.....	18
Литература.....	21
Приложение.....	22

Введение.

Цель курсовой работы

Целью предлагаемых учебно-практических занятий является изучение современных информационных технологий и инновационных разработок для сохранения исторических и культурных ценностей России на примере задачи исследования языка печатных источников XVIII – нач. XIX вв. Выполнение заданий позволяет на практике изучить особенности рассматриваемых текстов, проанализировать факторы, влияющие на эффективность их распознавания с помощью современных OCR-систем.

Задачи курсовой работы:

1. Подготовка материалов для выполнения задания: установка ПО, анализ и фиксация параметров ПО. Анализ характеристик источника.
2. Ввод и распознавание текстового фрагмента. Предварительная оценка эффективности работы OCR-системы.
3. Формирование шрифтовых эталонов фрагментов, используя технологию обучения. Оценка эффективности использования технологии распознавания, включающей шрифтовые эталоны.
4. Оценка эффективности использования технологии распознавания, включающей дополнительный словарь языка текста.
5. Корректурa текста и анализ лексики и типов ошибок.
6. Квантитативные исследования текста, формирование словников фрагментов, построение функций распределения частот. Выводы. Формирование отчета.

Основная часть.

Задача 1.

В качестве OCR-системы будет использоваться система ABBYY FineReader 10 Professional Edition.

Зафиксируем первоначальные настройки и характеристики системы.

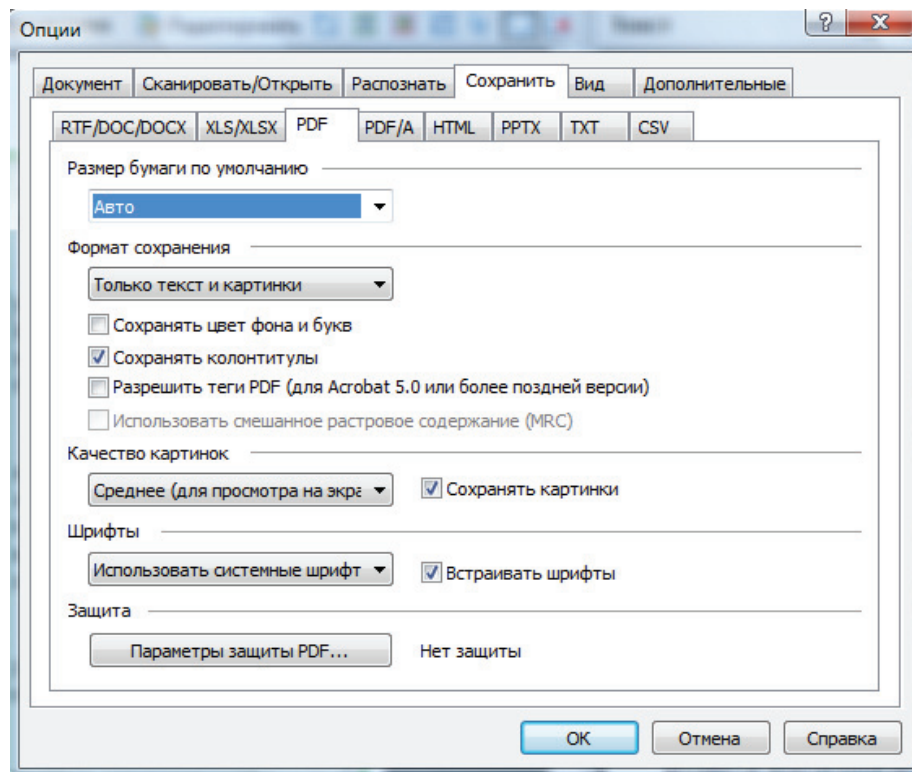


Рис.1. Первоначальные настройки системы.

Библиографические данные об источнике:

Лопухин, Иван Владимирович (1756-1816).

Изображение мечты равенства и буйной свободы, с пагубными их плодами. / Сочинение статского советника Ивана Лопухина. - Новое издание. - Москва : [Тип. Пономарева], 1794. - 30 с.; 8°.

Первоначально было опубликовано в кн. И. В. Лопухина "Излияние сердца" напечатанной в 1794 г. в Москве и Калуге.

Палеографические данные об источнике:

Формат и тип издания: старопечатная книга в электронном формате (.pdf)

Шрифт издания: переходная антиква, малоконтрастный

Данные о шрифтовых гарнитурах в Adobe Reader: Helvetica, Helvetica-Bold, Helvetica-Oblique, Times-Bold, Times-BoldItalic, Times-Italic, Times-Roman

Качество бумаги: хорошее, практически не пожелтевшая

Качество чернил: удовлетворительное, местами символы немного смазаны по краям

Следы времени: присутствуют (неравномерная толщина символов, в некоторых местах затёртости, некоторые символы практически не видны, слегка пожелтевшие страницы)

Задача 2.

Необходимо осуществить ввод фрагмента источника, распознавание с использованием OCR-системы (например ABBY FineReader).

Разработаем схему технологического процесса ввода текстовой информации.

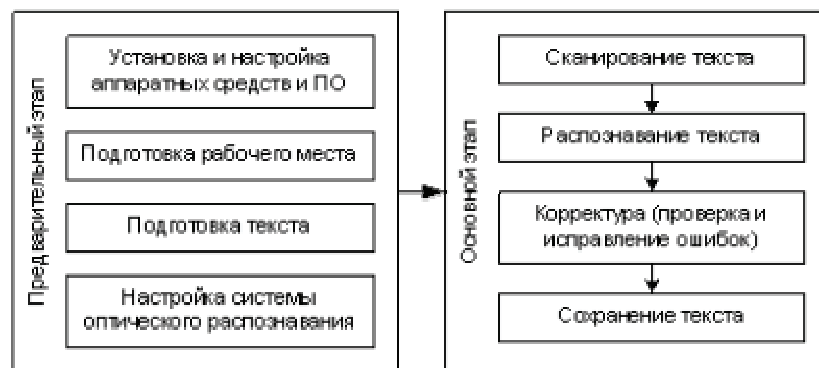


Рис. 2. Обобщенная схема технологического процесса ввода текста с помощью системы оптического распознавания.

Проведём распознавание документа, используя языки Русский и Русский (старая орфография).

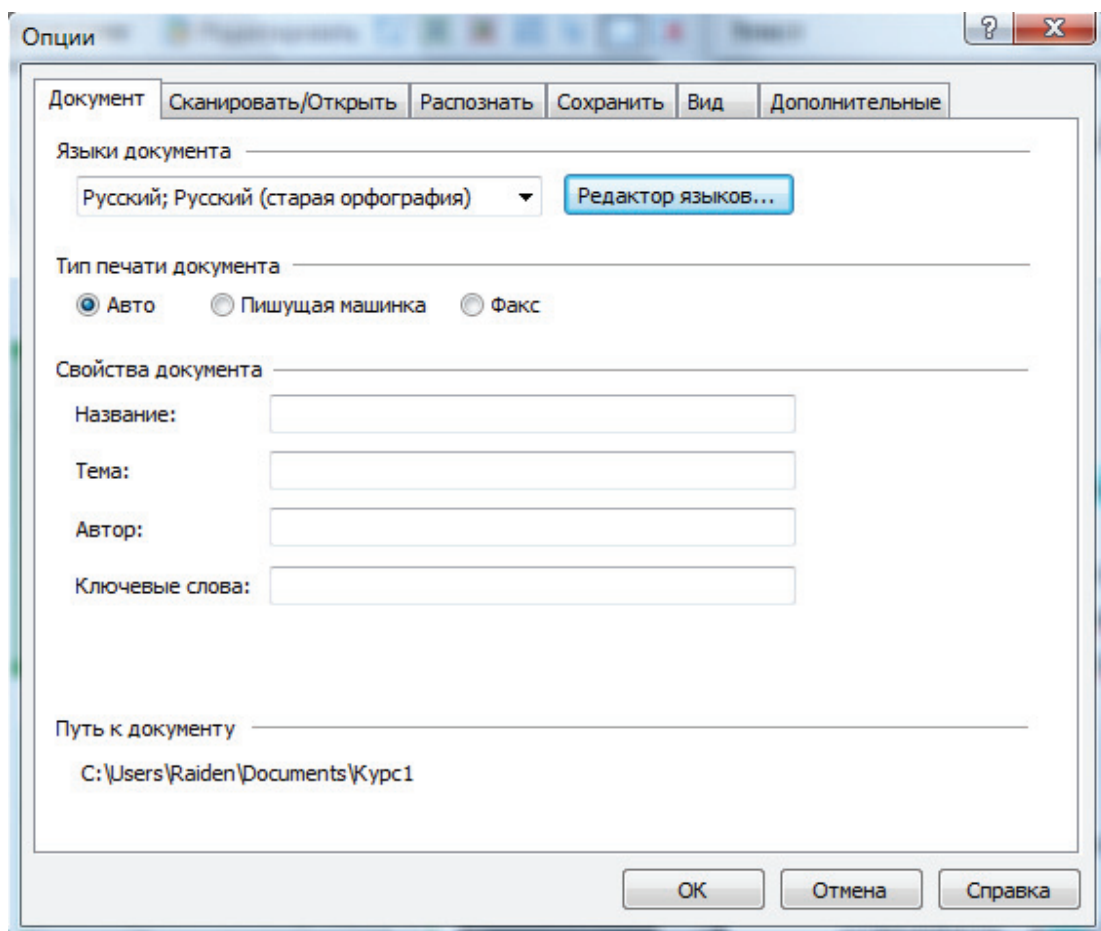


Рис. 3. Настройка языка распознавания.

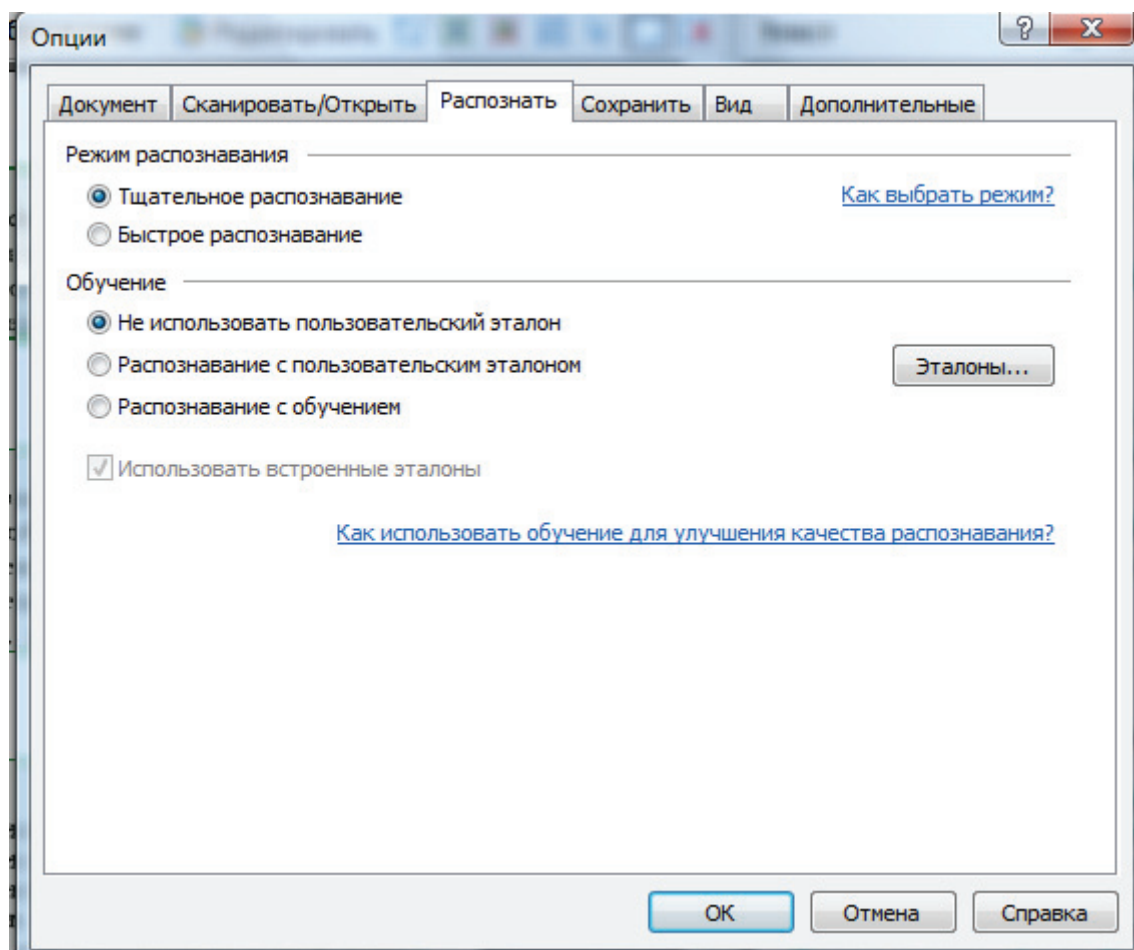


Рис. 4. Настройка режима распознавания.

Предварительная оценка эффективности работы OCR-системы включает анализ качества распознавания. Необходимо вычислить точность распознавания для каждой страницы. Для этого воспользуемся данными статистики и представим данные в таблице.

Табл. 1. Статистические данные распознавания.

Страница	Количество символов	Количество неуверенно распознанных символов	Точность распознавания
3	298	75	75 %
4	603	205	66 %
5	539	191	65 %
6	502	176	65 %
7	534	122	77 %
8	600	212	65 %
9	523	141	73 %
10	606	208	66 %
11	527	141	73 %
12	549	136	75 %
13	557	162	71 %
14	540	151	72 %
15	472	96	80 %
16	568	150	74 %
17	403	91	77 %
18	442	88	80 %
19	463	110	76 %
20	429	76	82 %
21	520	162	69 %
22	485	131	73 %
23	555	131	76 %
24	436	153	65 %
25	517	135	74 %
26	540	145	73 %
27	438	107	76 %
28	548	151	72 %

Временные затраты на распознавание: 29 сек.

Качество распознавания: невысокое, зафиксированная в среднем точность распознавания около 73 %.

Задача 3.

Далее необходимо проанализировать шрифты источника, построить алфавит, сформировать шрифтовые эталоны в OCR-системе, используя технологию обучения. Распознать текст, используя шрифтовые эталоны.

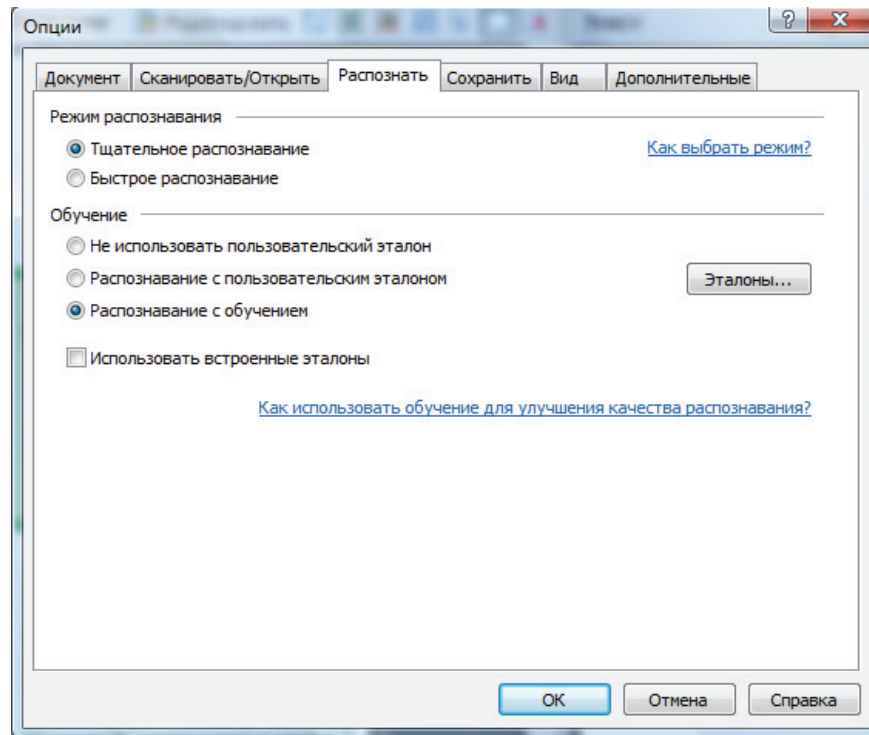


Рис. 5. Настройка режима распознавания с обучением.

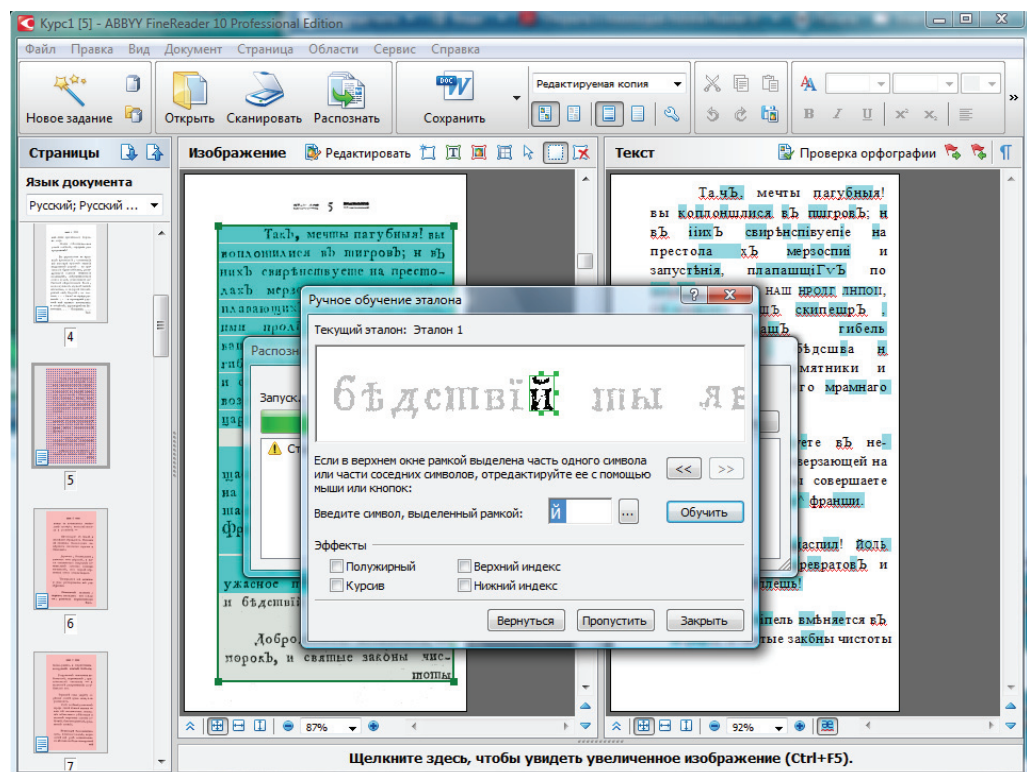


Рис. 6. Ручное обучение эталона.

Прохоров Артём, ИУ5-92

Ручное обучение проводится на трёх страницах текста (страницы 3-5).

В итоге получаем основную таблицу символов (алфавит) эталона:

! * , - . ; I A Б В Д Е Ж З И Й М Н О П Р С Т У Ф Ц Ч Ы а б в г д е ж з и й к л м н о п
р с т у ф х ц ч ш щ ь ы ю я ъ □

Среди символов алфавита буквы «ъ» и «□» имеют специфическое начертание (вышли из употребления).

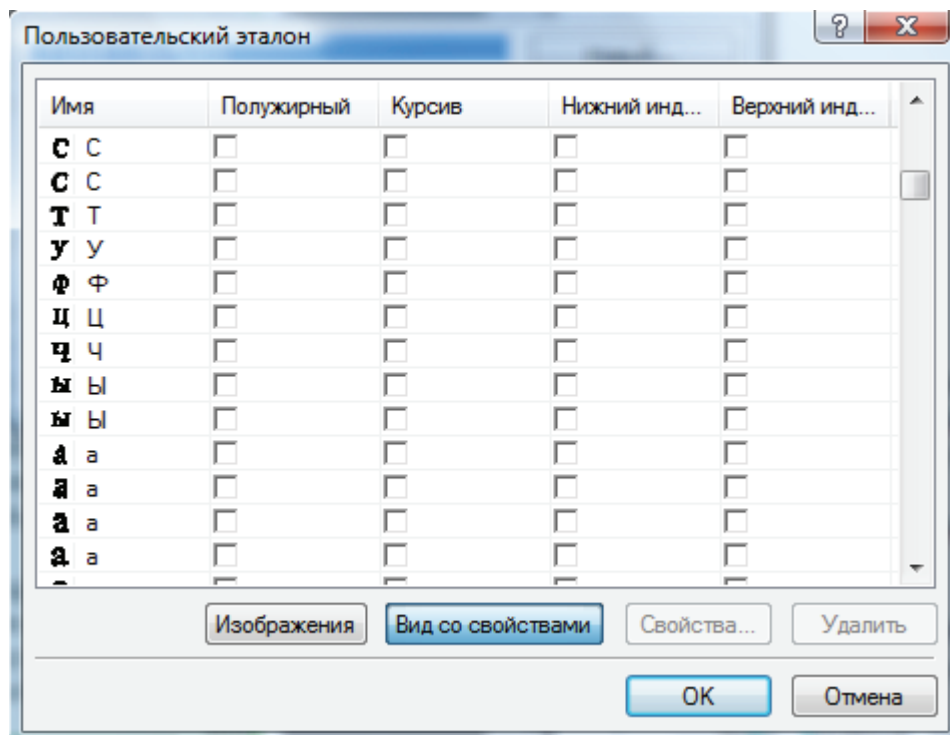


Рис. 7. Часть алфавита эталона.

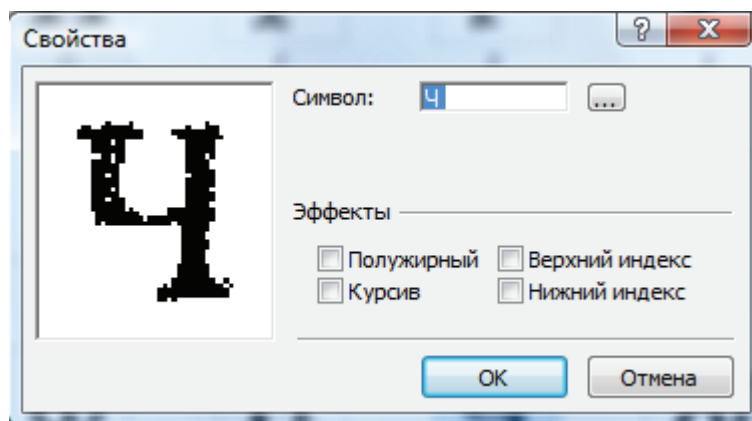


Рис. 8. Изображение символа.

Распознаем текст, используя полученный эталон.

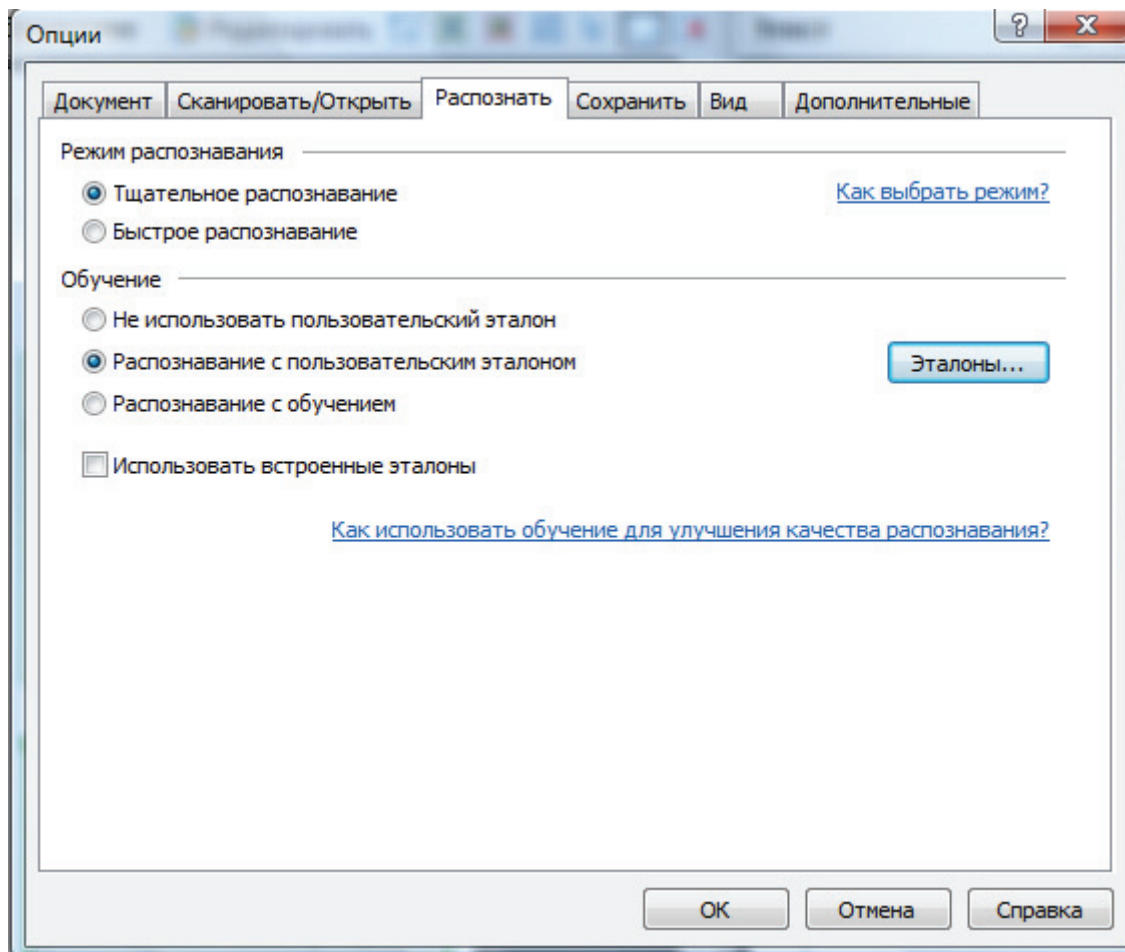


Рис. 9. Настройка режима распознавания с пользовательским эталоном.

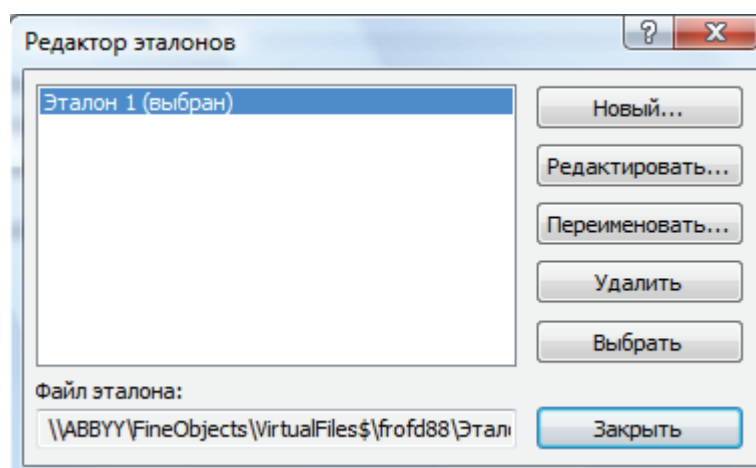


Рис. 10. Выбор эталона.

Оценим точность распознавания. Проанализируем временные затраты на обучение и распознавание с эталоном. Сформируем выводы об эффективности использования технологии распознавания с эталоном.

Табл. 2. Статистические данные распознавания с использованием пользовательского эталона.

Страница	Количество символов	Количество неуверенно распознанных символов	Точность распознавания
3	295	29	90 %
4	594	140	76 %
5	531	118	78 %
6	496	144	71 %
7	533	122	77 %
8	607	222	63 %
9	524	140	73 %
10	601	206	66 %
11	530	147	72 %
12	548	159	71 %
13	557	177	68 %
14	531	174	67 %
15	471	118	75 %
16	566	188	67 %
17	401	97	76 %
18	439	98	78 %
19	465	132	72 %
20	429	76	82 %
21	520	162	69 %
22	485	131	73 %
23	555	131	76 %
24	436	153	65 %
25	517	135	74 %
26	540	145	73 %
27	438	107	76 %
28	548	151	72 %

Временные затраты: на обучение – 1,5 часа, на распознавание – 31 сек.

Вывод: технология особенно эффективна на первых трёх распознаваемых страницах (на них производилось обучение), однако показывает хорошие результаты и на других. Требуется достаточно большое количество времени, так как необходимо провести ручное обучение эталона. Местами точность распознавания страниц исходя из количества неуверенно распознанных символов меньше, чем в предыдущем методе, однако по количеству фактических ошибок в символах и словах данный метод намного эффективнее предыдущего.

Задача 4.

Следует оценить эффективность технологии распознавания, включающей дополнительный словарь-спеллер языка XVIII в. Для этого подключим и настроим дополнительный пользовательский язык. Используется словник Словаря Академии Российской 1789-1794 гг.

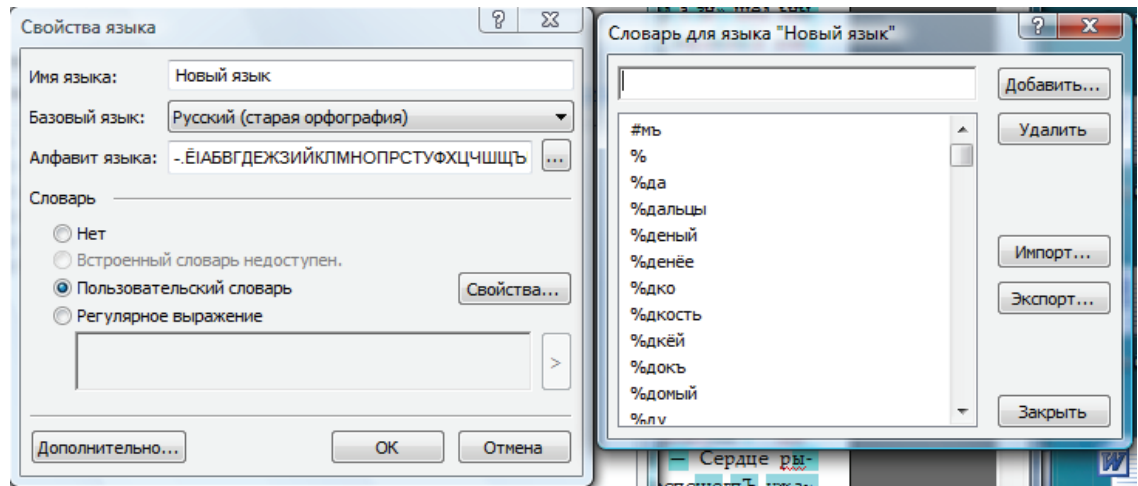


Рис. 11. Настройка языка «Новый язык».

Распознаем текст с использованием языка «Новый язык». Оценим точность распознавания. Проанализируем временные затраты на распознавание со словарем. Сформируем выводы об эффективности использования технологии распознавания с дополнительным словарем-спеллером.

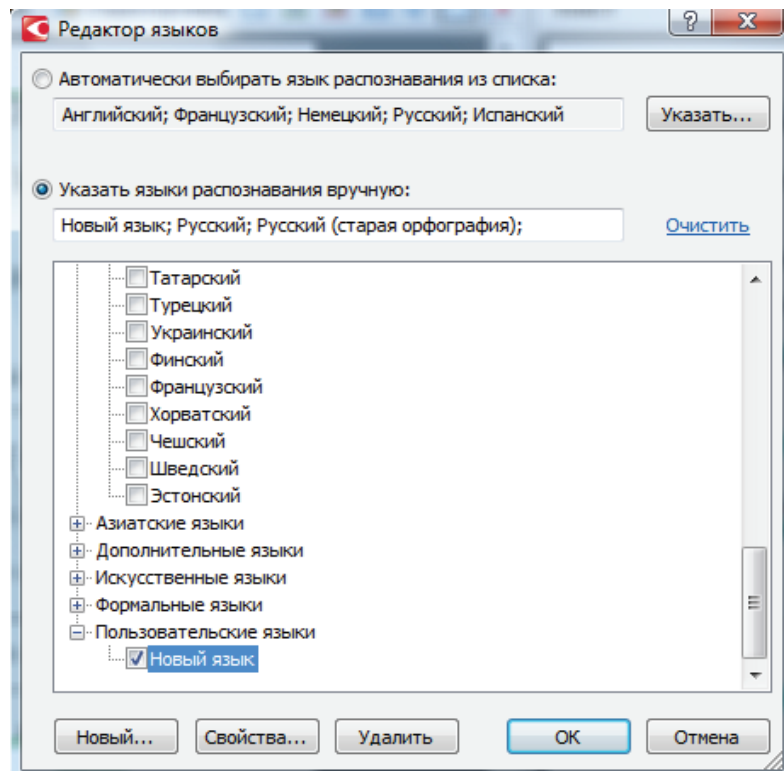


Рис. 12. Выбор языка «Новый язык».

Прохоров Артём, ИУ5-92

Табл. 3. Статистические данные распознавания с использованием пользовательского языка.

Страница	Количество символов	Количество неуверенно распознанных символов	Точность распознавания
3	300	68	77 %
4	591	206	65 %
5	539	198	63 %
6	502	175	65 %
7	534	120	78 %
8	607	211	65 %
9	521	146	72 %
10	608	212	65 %
11	526	138	74 %
12	553	144	74 %
13	561	183	67 %
14	542	150	72 %
15	469	94	80 %
16	568	162	71 %
17	406	93	77 %
18	441	88	80 %
19	463	126	73 %
20	424	73	83 %
21	512	158	69 %
22	483	136	72 %
23	551	121	78 %
24	439	146	67 %
25	514	142	72 %
26	535	149	72 %
27	439	116	74 %
28	553	156	72 %

Временные затраты на распознавание – 29 сек.

Вывод: технология показывает хорошие результаты, при этом точность распознавания на некоторых страницах ниже, а на некоторых выше, чем в предыдущих методах. Однако, судя по количеству фактических ошибок, этот метод также гораздо менее эффективен, чем метод распознавания с использованием пользовательского эталона

Задача 5.

Так как метод распознавания с использованием пользовательского эталона оказался наиболее эффективным по количеству фактических ошибок, будем корректировать текст именно на его основе. Сначала проанализируем фактическую точность распознавания символов в данном методе.

Табл. 4. Статистические данные распознавания символов.

Страница	Количество символов	Количество ошибок (в символах)	Точность распознавания символов
3	295	1	99 %
4	594	21	96 %
5	531	19	96 %
6	496	25	95 %
7	533	25	95 %
8	606	31	95 %
9	524	37	93 %
10	601	44	93 %
11	530	21	96 %
12	547	27	95 %
13	558	36	94 %
14	531	42	92 %
15	471	18	96 %
16	569	42	93 %
17	400	17	96 %
18	440	17	96 %
19	467	32	93 %
20	426	11	97 %
21	510	40	92 %
22	482	20	96 %
23	553	35	94 %
24	438	29	93 %
25	513	27	95 %
26	528	28	95 %
27	441	19	96 %
28	551	34	94 %

Как мы видим, фактическая точность распознавания символов достаточно высока (не ниже 92 %).

Теперь проанализируем лексический состав источника и рассмотрим так называемые несловарные слова, которые проверяются с помощью словаря-спеллера.

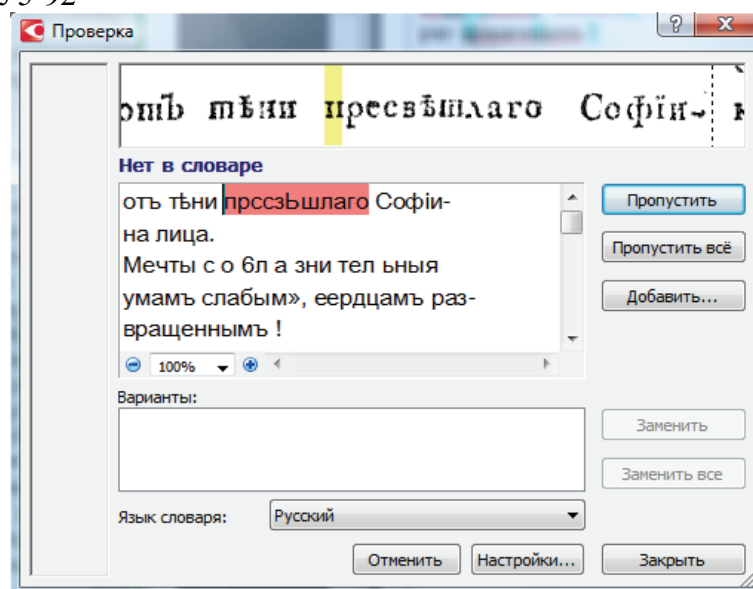


Рис. 13. Проверка несловарных слов.

Табл. 5. Статистические данные распознавания слов.

Страница	Количество слов	Количество ошибочных слов	Точность распознавания слов	Количество проверяемых несловарных слов	Количество ошибочных несловарных слов	Точность распознавания несловарных слов
3	35	1	97 %	11	1	91 %
4	65	15	77 %	27	4	85 %
5	73	13	82 %	15	3	80 %
6	58	12	79 %	15	5	66 %
7	62	16	74 %	23	6	74 %
8	75	24	68 %	24	14	42 %
9	73	19	74 %	23	8	65 %
10	81	22	73 %	17	13	24 %
11	62	15	76 %	15	11	27 %
12	79	20	75 %	17	11	35 %
13	85	21	75 %	18	9	50 %
14	70	24	66 %	22	11	50 %
15	62	13	79 %	20	7	65 %
16	70	25	74 %	19	7	63 %
17	46	10	78 %	14	4	71 %
18	52	10	81 %	17	5	71 %
19	61	18	70 %	14	8	43 %
20	53	6	89 %	13	6	54 %
21	61	21	66 %	24	12	50 %
22	60	12	80 %	22	10	55 %
23	74	18	76 %	19	7	63 %
24	59	13	78 %	18	7	61 %
25	71	18	75 %	16	6	62 %
26	72	18	75 %	20	9	55 %
27	57	12	79 %	14	6	57 %
28	72	19	74 %	23	15	35 %

Проиллюстрируем полученные данные графиком.

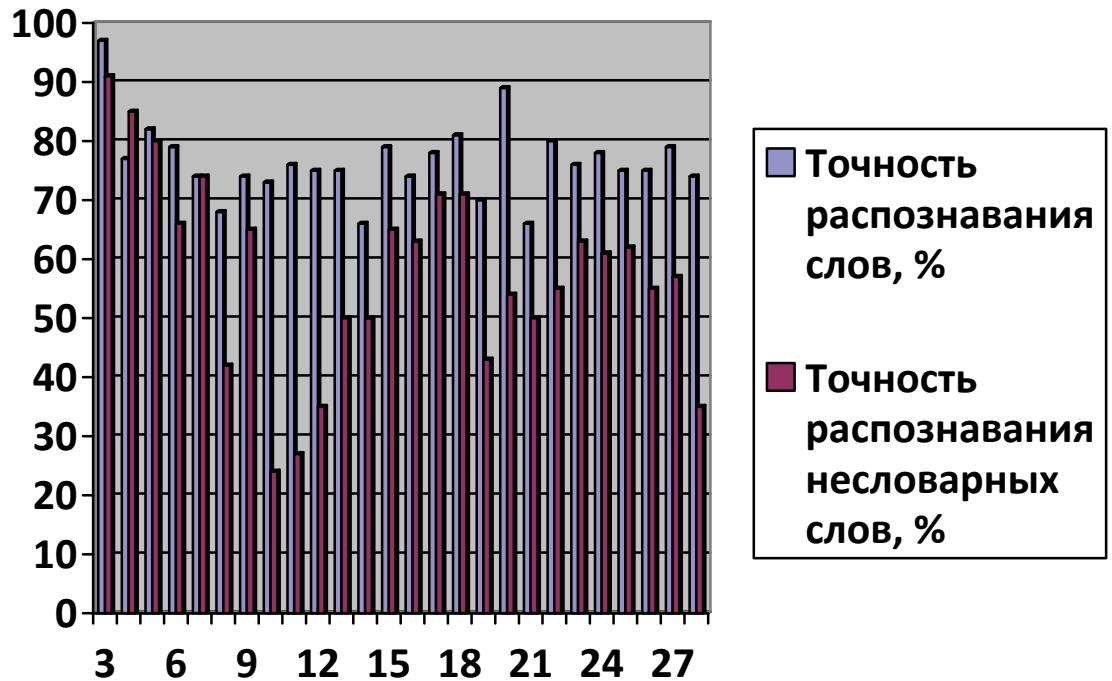


Рис. 14. Диаграмма - точность распознавания словарных и несловарных слов на каждой странице.

Мы видим, что точность распознавания несловарных слов в целом ниже, чем общая точность распознавания слов на странице. Теперь анализируем наиболее часто встречающиеся ошибки в символах и причины их возникновения.

Табл. 6. Типы ошибок.

Тип ошибки	Пример		Причина возникновения	Количество в тексте
	Ошибка	Исправление		
п-ш	нещасныхЪ	нещасныхЪ	Сходство графем символов	73
е-с	ссердсѣЪ	сердѣЪ	Сходство графем символов	37
полужирный шрифт	мечпы	мечпы	Неравномерность толщины штриха в оригинале	22
в-з	пресзѣплаго	пресвѣплаго	Сходство графем символов	9
н-я	сыяовЪ	сыновЪ	Сходство графем символов	12
п - ip (pi)	рыдаеіпЪ	рыдаепЪ	Сходство графем символов	8
я - л	суевѣрїл	суевѣрїя	Сходство графем символов	14
й - н	про-свѣпипеле н	про-свѣпипелей	Сходство графем символов	23
ѣ - ъ	книгЪ	книгѣ	Сходство графем символов	17

Вывод: произведён анализ наиболее часто встречающихся ошибок, а также их количества и причин. Приведены данные о точности распознавания с использованием метода распознавания с помощью пользовательского эталона. На основе этого метода произведена корректировка текста. Пример корректировки приведён в Приложении к данной Записке.

Задача 6.

Проведём квантитативные исследования. Сначала построим частотный словник всего откорректированного текста. Для этих целей используем программу INTERLEX, дополнив собственными вычислениями полученный в программе словник.

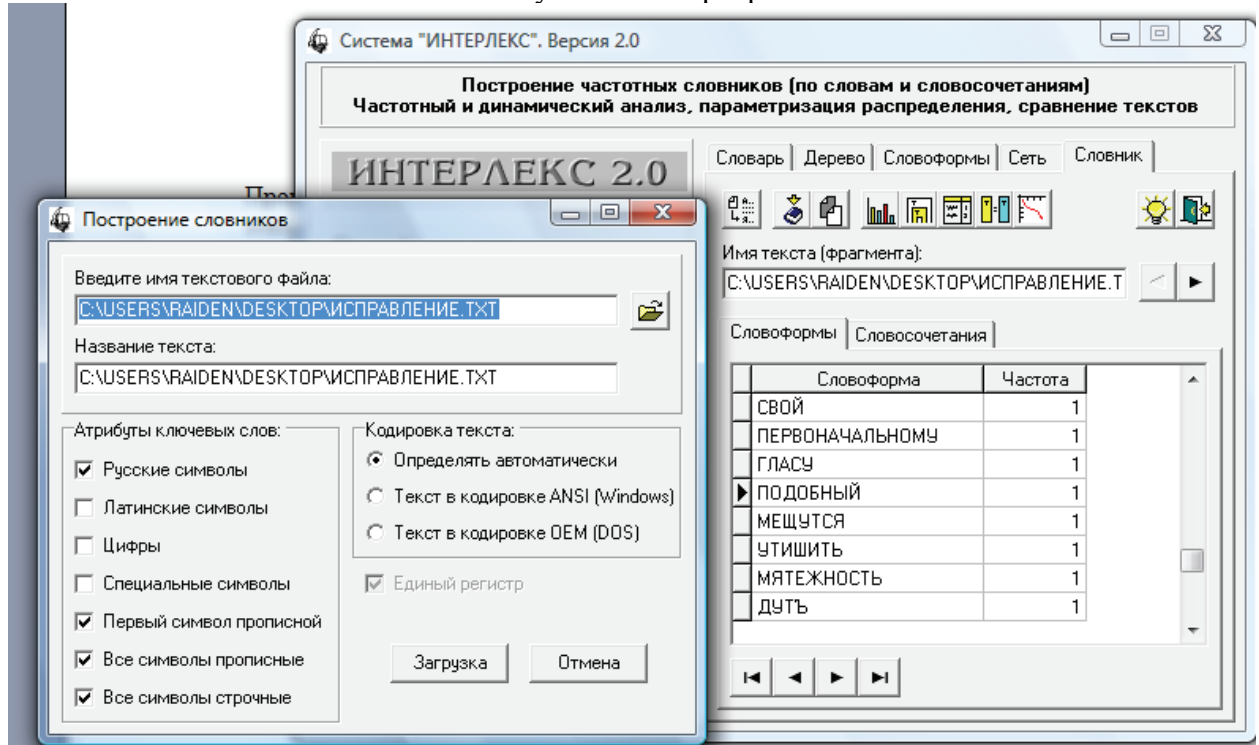


Рис. 15. Построение словника в программе INTERLEX.

Приведём фрагмент полученного частотного словника (весь словник приведён в папке Частота_Прохоров).

Табл. 7. Фрагмент частотного словника текста.

#	Словоформа	i	i/K	r	r^{-1}	$0.1 r^{-1}$
1	И	78	0,102902	1	1	0,1
2	ВЪ	41	0,05409	2	0,5	0,05
3	НА	28	0,036939	3	0,333333	0,033333
4	ЕГО	28	0,036939	3	0,333333	0,033333
5	БЫ	21	0,027704	4	0,25	0,025
6	ВО	16	0,021108	5	0,2	0,02
7	ИХЪ	14	0,01847	6	0,166667	0,016667
8	НЕ	13	0,01715	7	0,142857	0,014286
9	ЕСТЬ	11	0,014512	8	0,125	0,0125
10	ОТЪ	10	0,013193	9	0,111111	0,011111

Прохоров Артём, ИУ5-92

11	КЪ	9	0,011873	10	0,1	0,01
12	НО	9	0,011873	10	0,1	0,01
13	ЕЯ	8	0,010554	11	0,090909	0,009091
14	ИЗЪ	8	0,010554	11	0,090909	0,009091
15	ВСЕ	7	0,009235	12	0,083333	0,008333
16	ВЫ	7	0,009235	12	0,083333	0,008333
17	СЕБЯ	7	0,009235	12	0,083333	0,008333
18	СВОИХЪ	7	0,009235	12	0,083333	0,008333
19	ИЛИ	6	0,007916	13	0,076923	0,007692
20	ВСЯ	6	0,007916	13	0,076923	0,007692
21	БЫТЬ	6	0,007916	13	0,076923	0,007692
22	ТОЛЬКО	6	0,007916	13	0,076923	0,007692
23	УЖЕ	6	0,007916	13	0,076923	0,007692
24	СВОБОДЫ	6	0,007916	13	0,076923	0,007692
25	ДЛЯ	5	0,006596	14	0,071429	0,007143
26	ЯКО	5	0,006596	14	0,071429	0,007143
27	ПО	4	0,005277	15	0,066667	0,006667
28	ЕДИНЫЙ	4	0,005277	15	0,066667	0,006667
29	ЖЕ	4	0,005277	15	0,066667	0,006667
30	СВЯЩЕННОЮ	4	0,005277	15	0,066667	0,006667
31	СУЩЕСТВО	4	0,005277	15	0,066667	0,006667
32	КОЛЬ	4	0,005277	15	0,066667	0,006667
33	МОЖЕТЪ	4	0,005277	15	0,066667	0,006667
34	ОБРАЗЪ	4	0,005277	15	0,066667	0,006667
35	ДУШИ	4	0,005277	15	0,066667	0,006667
36	ЕЖЕЛИ	4	0,005277	15	0,066667	0,006667
37	О	4	0,005277	15	0,066667	0,006667
38	КО	4	0,005277	15	0,066667	0,006667
39	БЫЛИ	4	0,005277	15	0,066667	0,006667
40	МЕЧТЫ	4	0,005277	15	0,066667	0,006667
41	ИБО	3	0,003958	16	0,0625	0,00625
42	СВОЕ	3	0,003958	16	0,0625	0,00625
43	ЧТО	3	0,003958	16	0,0625	0,00625
44	КАКЪ	3	0,003958	16	0,0625	0,00625
45	ТО	3	0,003958	16	0,0625	0,00625
46	ЛЮБВИ	3	0,003958	16	0,0625	0,00625
47	А	3	0,003958	16	0,0625	0,00625
48	СИЛЫ	3	0,003958	16	0,0625	0,00625
49	ЗЛАТОЙ	3	0,003958	16	0,0625	0,00625

Прохоров Артём, ИУ5-92

50	СВОБОДА	3	0,003958	16	0,0625	0,00625
...	...	...	...	...	...	...

Построим гистограмму функции распределения абсолютной и относительной частот слов фрагмента текста.

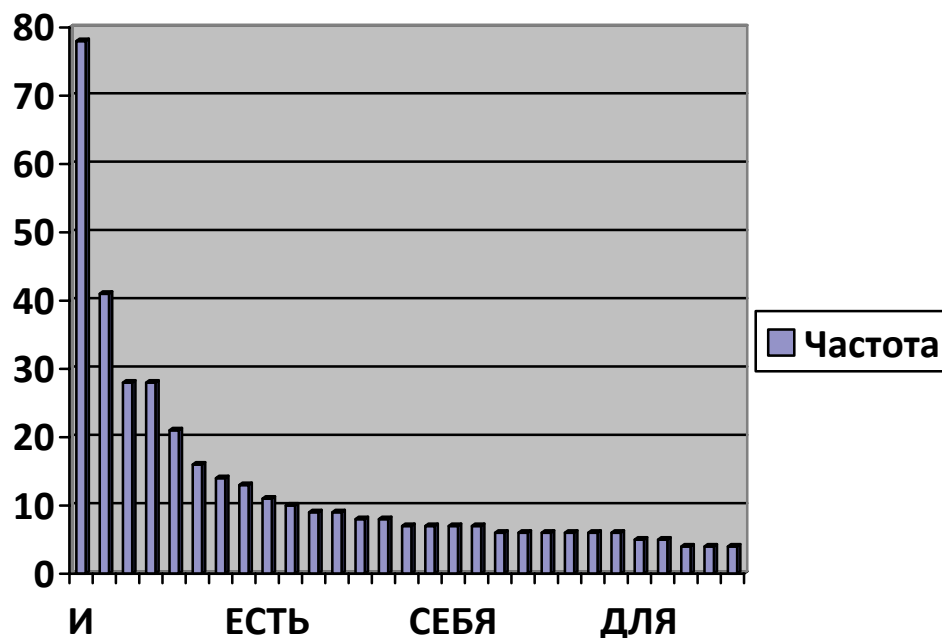


Рис. 16. Гистограмма функции распределения абсолютной и относительной частот слов фрагмента текста.

Вывод: мы провели распознавание и анализ, корректировку текста на основе метода распознавания с помощью пользовательского эталона. Основным недостатком данного метода являются временные затраты, так как они достаточно велики из-за необходимости ручного обучения эталона. Основным его достоинством является наименьшее (среди исследуемых методов) количество фактических ошибок в тексте. Для данного источника произведена корректировка, проведены все требуемые исследования, в том числе и количественные. Результаты корректировки и исследований приведены на диске вместе с электронным вариантом данной Записки.

Литература.

1). *Филиппович Анна*. Исследование эффективности систем оптического распознавания текстов. // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. Выпуск 7 / Сост. и ред. Ю.Н. Филипповича. – М. Изд-во ООО «Эликс+» 2005. – С. 272-297.

2). *Филиппович Анна*. Курс лекций по дисциплине «Лингвистическое обеспечение АСОИУ».

3). Знакомство с программой ABBYY FineReader. <http://finereader.helpmax.net/ru/>

Приложение.

Исходная страница текста.



ИЗОБРАЖЕНІЕ
МЕЧТЫ РАВЕНСТВА
и
БУЙНОЙ СВОБОДЫ,
съ
пагубными ихъ плодами.

Равенство! свобода буйная!
мечты, порожденныя чадомъ
тусклаго свѣшильника лжему-
дрія; распложденныя безумными
писаніями нечестивыхъ патей
философскаго имени, адскимъ
пламенемъ стремящихся опира-
щать взоръ человѣческій даже
отъ

Распознанная страница текста.

ИЗОБРАЖЕНІЕ
МЕЧТЫ РАВЕНСТВА И
БУЙНОЙ СВОБОДЫ, съ
пагубными ихъ плодами.

Равенство! свобода буйная !
мечты, порожденныя чадомъ
тусклаго свѣтильника лжему- дрія:
расположенныя безумными
писаніями Нечестивыхъ татей
философскаго имени, адскимъ
пламенемъ стремящихся отвращать
взоръ чловѣческій даже

Прохоров Артём, ИУ5-92

Откорректированная страница текста.

ИЗОБРАЖЕНІЕ
МЕЧТЫ РАВЕНСТВА
И
БУЙНОЙ СВОБОДЫ,
СЪ
пагубными ихъ плодами.

Равенство! свобода буйная!
мечты, порожденныя чадомъ
пусклаго свѣшльника лжему-
дрія; распложенныя безумными
писаніями нечестивыхъ патей
Философскаго имени адскимъ
пламенемъ стремящихся отвра-
щать взоръ человѣческій даже
отъ