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 ИУ5-95
Левкин Константин Сергеевич.

Шифр:95_06

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

Москва, 2010 г.

Содержание

1. 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. Корректура текста и анализ лексики и типов ошибок.
6. Квантитативные исследования текста, формирование словников фрагментов, построение функций распределения частот. Выводы. Формирование отчета.

2. Основная часть.

Задача 1

Установить ПО, используемое для дальнейшего распознавания текста – OCR-систему (например FineReader). Перед вводом текста необходимо зафиксировать исходные характеристики и настройки программного и аппаратного обеспечения, используемого для ввода.

Переписать фрагмент источника (изображения страниц). Указать характеристики изображений страниц текста.

Указать библиографические данные источника. Проанализировать палеографические характеристики источника: указать тип и формат издания, используемые шрифтовые гарнитуры, качество бумаги, чернил, следы времени и т.п.

++

Аппаратное и программное обеспечение:

- Процессор – Intel Core 2 Duo P8400 2,26GHz 2,27 GHz;
- ОЗУ – 4 Гб;
- ОС – Microsoft Windows 7;
- Используемое ПО – ABBYY FineReader 9.0 Professional Edition.

Исходные настройки ABBYY FineReader 9.0 PE:

- Режим распознавания – Тщательное распознавание;
- Обучение – Не использовать пользовательский эталон;
- Тип области распознавания – Текст.

Характеристики изображения:

- Ширина*высота: 1265*2039 pt;
- Разрешение: 300 dpi;
- Тип изображения: черно-белое PDF;
- Качество сканированного изображения – в основном хорошее и несколько страниц среднего качества (тёмные пятна от пометок, сдвигов и инородных загрязнителей).

Библиографические и палеографические характеристики источника:

- ЭБ РГБ – Старопечатные книги;
- «Главы к уставу о полевой службе»;
- Санкт-Петербург: Тип. Воен. коллегии, 1792г., 63с.;
- Хранение: МК ВК-8°/ 92-Г;
- Следы времени – несущественные.

Задача 2

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

Разработать схему технологического процесса ввода текстовой информации, представить ее описание (этапы, процедуры и операции).

Предварительная оценка эффективности работы OCR-системы включает анализ качества распознавания. Необходимо вычислить точность распознавания для каждой страницы. Представить данные в таблице. Проанализировать временные затраты и качество распознавания, настроить опции OCR-системы для оптимальной работы.

++

Схема технологического процесса ввода текстовой информации:

1. Запустить программу ABBYY FineReader 9.0 PE.
2. Создать новый документ FineReader, сохранить его.

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

3. Открыть отсканированные страницы текста.

Для этого необходимо нажать кнопку «Открыть» или Файл → Открыть PDF/Изображение.. (Ctrl+O). После этого возникает окно с изображением страницы, окно «текст» и окно укрупненного изображения.

4. Отобрать необходимые страницы, удалив ненужные.
5. Распознать текст.

Для распознавания текста необходимо нажать кнопку «Распознать документ» или Документ → Распознать документ (Ctrl+Shift+R). После распознавания в окне «текст» появится сам распознанный текст.

Для распознавания с обученным эталоном в опциях распознавания выбрать пункт «Распознавание с пользовательским эталоном».

Для подключения словаря в опциях Дополнительные указать папку, содержащую словарь.

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

Для этого необходимо нажать кнопку «Сохранить» или Файл → Сохранить как → Документ Microsoft Word. После этого будет запущена программа Microsoft Word и текст появится на экране. Его следует сохранить в формате *.docx, для этого в меню «Файл» необходимо выбрать «Сохранить».

Предварительная оценка эффективности работы OCR-системы:

Таблица 1. Статистические данные распознавания

Страница	Количество символов	Количество неуверенно распознанных символов	Точность распознавания %
7	749	121	83,85
8	926	210	77,32
9	838	193	76,97
10	949	174	81,66
11	874	122	86,04
12	969	186	80,80
13	980	107	89,08
14	1003	84	91,63
15	959	178	81,44
16	908	194	78,63
17	1036	297	71,33
18	830	172	79,28
19	872	95	89,11
20	745	164	77,99
21	843	165	80,43
22	860	76	91,16
23	977	94	90,38
24	919	255	72,25
25	895	192	78,55
26	800	107	86,63
27	919	112	87,81
28	776	195	74,87
29	898	142	84,19
30	790	113	85,70
31	905	149	83,54

Итог:

В качестве временных затрат будем рассматривать затраты непосредственно на само распознавание, т. к. время подготовительного процесса зависит от конкретного пользователя и ПО.

Таким образом, на распознавание 25 страниц потребовалась 1 минута 15 секунд. Средняя точность распознавания = 82,42%. Данная точность обусловлена отсутствием некоторых старорусских букв в современном русском языке, используемом при распознавании, а также не отличным качеством отсканированного изображения.

Задача 3

Проанализировать шрифты источника, построить алфавит, сформировать шрифтовые эталоны в OCR-системе, используя технологию обучения. Распознать текст, используя шрифтовые эталоны. Следует обратить внимание на *неуверенно распознанные символы*, имеющие специфическое написание или вышедшие из употребления буквы.

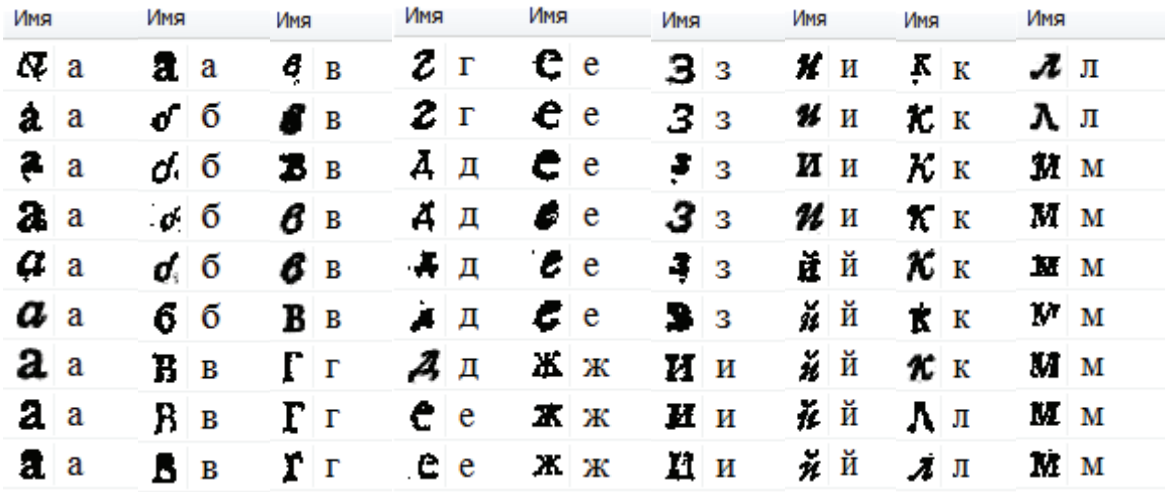
Представить полученные эталоны, описать шрифты, основную таблицу символов (алфавит). Оценить точность распознавания. Проанализировать временные затраты на обучение и распознавание с эталоном. Сформировать выводы об эффективности использования технологии распознавания с эталоном.

++

Шрифт, используемый в данном источнике:

- По ГОСТ 3489.1-71: малоконтрастный шрифт;
- По начертанию:
  - прямой;
  - курсивный;
- По исторической классификации: переходная антиква;

Пользовательский шрифтовой эталон:



Алфавит:

1, 2, 3, 4, 5, 6, 7, 8, 9, 0

а, б, в, г, д, е, ж, з, и, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ъ, ы, ь, ю, я

А, Б, В, Г, Д, Е, Ж, З, И, К, Л, М, Н, О, П, Р, С, Т, У, Ф, Х, Ц, Ч, Ш, Щ, Ъ, Ы, Ь, Ю, Я

Эффективность распознавания с эталоном:

Таблица 2. Статистические данные распознавания с использованием пользовательских эталонов

Страница	Количество символов	Количество неуверенно распознанных символов	Точность распознавания
7	749	97	87,05
8	926	159	82,83
9	838	196	76,61
10	949	160	83,14
11	874	98	88,79
12	969	173	82,15
13	980	105	89,29
14	1003	86	91,43
15	959	161	83,21
16	908	195	78,52
17	1036	269	74,03
18	830	156	81,20
19	872	87	90,02
20	745	143	80,81
21	843	148	82,44
22	860	64	92,56
23	977	58	94,06
24	919	228	75,19
25	895	193	78,44
26	800	75	90,63
27	919	87	90,53
28	776	174	77,58
29	898	136	84,86
30	790	96	87,85
31	905	133	85,30

Итог:

Временные затраты на обучение эталона составили около 16 минут (3 страницы), но на распознавание 25 страниц в режиме Распознавание с пользовательским эталоном ушло порядка 40 секунд, т. е. почти вдвое быстрее обычного распознавания.

Что же касается средней точности распознавания – она повысилась, но не намного, и составила 84,34%. Существенное повышения точности распознавания можно заметить на обучаемых страницах и на страницах без курсивного шрифта.

Можно сделать следующие выводы. Данная технология эффективна при наиболее обученном эталоне (т. е. когда временные рамки не сильно ограничены), при необходимости распознавания огромных объёмов текста (т. е. время обучения эталона мало относительно времени распознавания текста).

Задача 4

Оценить эффективность технологии распознавания, включающей дополнительный словарь-спеллер языка XVIII в. (используя словник Словаря Академии Российской 1789-1794 гг.).

Оценить точность распознавания. Проанализировать временные затраты на распознавание со словарем. Сформировать выводы об эффективности использования технологии распознавания с дополнительным словарем-спеллером.

++

Эффективность распознавания со Словарём Академии Российской:

Таблица 3. Статистические данные распознавания с использованием дополнительного словаря-спеллера XVIII в.

Страница	Количество символов	Количество неуверенно распознанных символов	Точность распознавания
7	749	89	88,12
8	926	143	84,56
9	838	170	79,71
10	949	142	85,04
11	874	81	90,73
12	969	156	83,90
13	980	89	90,92
14	1003	80	92,02
15	959	149	84,46
16	908	177	80,51
17	1036	224	78,38
18	830	139	83,25
19	872	78	91,06
20	745	135	81,88
21	843	132	84,34
22	860	67	92,21
23	977	62	93,65
24	919	203	77,91
25	895	180	79,89
26	800	73	90,88
27	919	69	92,49
28	776	136	82,47
29	898	114	87,31
30	790	82	89,62
31	905	101	88,84

Итог:

Распознавание со словарём заняло порядка 45 секунд, что ненамного превышает время распознавания с пользовательским эталоном.

Средняя точность распознавания еще немного повысилась и составила 86,17%. Данный метод улучшил качество распознавания за счет определения старинных символов (например: ѣ, ї). Следовательно, можно сделать вывод, что распознавание со словарём-спеллером будет наиболее эффективно при распознавании текстов с большим количеством символов, не используемых в современном алфавите.

Задача 5

Откорректировать текст источника, проанализировать проверяемые слова и символы. Рассмотреть *неуверенно распознанные и нераспознанные символы*, сделать предположение о причинах сложности распознавания, привести примеры, посчитать количественные характеристики, построить иллюстрирующие графики. Символы, точность распознавания которых очень низка следует представить отдельной таблицей и проиллюстрировать.

++

Анализ основных ошибок:

При корректировки текста источника был выявлен ряд наиболее ошибочно распознающихся символов.

Самые большие проблемы возникли при распознавании символов «т» и «ш». Из-за наличия довольно больших засечек буквы «т» и «ш» выглядят почти одинаково, а из-за того, что «т» встречается значительно чаще «ш», данная замена более распространена. ***Пример:*** «путки» вместо «пушки», «больтей» вместо «большей», «рапоршуют» вместо «рапортуют», «под предводишельством» вместо «под предводительством».

Следующая распространённая ошибочная замена «е» и «с». Из-за похожего написания данных символов и особенно из-за не очень высокого качества печати возникала данная ошибка (особенно при курсивном начертании шрифта). ***Пример:*** «чаеовыхъ» вместо «часовыхъ», «бееконечно» вместо «бесконечно».

Ещё одной частой ошибкой является распознавание «п» как «л» из-за малого различия в начертании и среднего качества сканирования источника. ***Пример:*** «фпангамъ» вместо «флангамъ», «в попкахъ» вместо «в полкахъ».

Также довольно часто встречается замена «я» на «л» и «н» на «я». Это связано с похожестью данных символов не достаточно хорошего качества печати в 18 веке. ***Пример:*** «ружьл» вместо «ружья», «на смеяе» вместо «на смене».

Чуть менее часто встречается ошибочное распознавание лигатур: «лт» вместо «ми», «шьі» вместо «ты». Это происходит из-за малого расстояния между соседними символами в слове. ***Пример:*** «салтм» вместо «самим», «пехошьі» вместо «пехоты».

Периодически встречалось нечёткое или неверное распознавание символов вышедших из употребления «ѣ» и «ї». Это случалось при распознавании без подключенного чловаря CAP. ***Пример:*** «батал’оне» вместо «баталїонѣ».

Количественные характеристики ошибок распознавания:

Таблица 4. Статистические данные распознавания символа «т»

Страница	Количество символов	Количество ошибок (в символах)	Точность распознавания символов
7	35	6	82,86
8	37	9	75,68
9	44	11	75,00
10	41	8	80,49
11	52	4	92,31
12	57	5	91,23
13	43	3	93,02
14	46	7	84,78
15	35	2	94,29
16	56	6	89,29
17	48	2	95,83
18	43	4	90,70
19	29	0	100,00
20	41	3	92,68
21	41	7	82,93

22	37	3	91,89
23	32	4	87,50
24	36	8	77,78
25	53	11	79,25
26	41	7	82,93
27	56	7	87,50
28	44	1	97,73
29	49	2	95,92
30	42	2	95,24
31	38	1	97,37

Таблица 5. Статистические данные распознавания символа «е»

Страница	Количество символов	Количество ошибок (в символах)	Точность распознавания символов
7	46	3	93,48
8	53	2	96,23
9	46	5	89,13
10	56	5	91,07
11	51	4	92,16
12	60	7	88,33
13	66	6	90,91
14	69	6	91,30
15	58	5	91,38
16	49	1	97,96
17	52	4	92,31
18	53	8	84,91
19	64	7	89,06
20	47	2	95,74
21	62	2	96,77
22	55	6	89,09
23	59	5	91,53
24	70	9	87,14
25	54	5	90,74
26	49	0	100,00
27	47	0	100,00
28	54	1	98,15
29	60	4	93,33
30	54	3	94,44
31	31	0	100,00

Таблица 6. Статистические данные распознавания символа «п»

Страница	Количество символов	Количество ошибок (в символах)	Точность распознавания символов
7	18	2	88,89
8	30	3	90,00
9	21	2	90,48
10	32	1	96,88
11	28	5	82,14
12	23	4	82,61
13	33	4	87,88
14	36	7	80,56
15	32	5	84,38
16	31	3	90,32
17	27	0	100,00
18	16	0	100,00

19	16	0	100,00
20	22	1	95,45
21	15	1	93,33
22	23	2	91,30
23	20	1	95,00
24	33	2	93,94
25	28	5	82,14
26	18	4	77,78
27	12	2	83,33
28	19	5	73,68
29	25	3	88,00
30	19	3	84,21
31	14	1	92,86

Таблица 7. Статистические данные распознавания символа «ѣ» (с пользовательским эталоном)

Страница	Количество символов	Количество ошибок (в символах)	Точность распознавания символов
7	12	11	8,33
8	22	20	9,09
9	17	17	0,00
10	17	16	5,88
11	14	14	0,00
12	16	16	0,00
13	11	9	18,18
14	17	14	17,65
15	15	14	6,67
16	12	10	16,67
17	15	15	0,00
18	10	10	0,00
19	4	4	0,00
20	11	11	0,00
21	8	7	12,50
22	9	7	22,22
23	7	7	0,00
24	6	4	33,33
25	8	8	0,00
26	11	10	9,09
27	10	10	0,00
28	8	7	12,50
29	5	3	40,00
30	12	9	25,00
31	5	4	20,00

Таблица 8. Статистические данные распознавания символа «ѣ» (со словарём CAP)

Страница	Количество символов	Количество ошибок (в символах)	Точность распознавания символов
7	12	4	66,67
8	22	13	40,91
9	17	7	58,82
10	17	6	64,71
11	14	5	64,29
12	16	9	43,75
13	11	8	27,27

14	17	6	64,71
15	15	10	33,33
16	12	8	33,33
17	15	4	73,33
18	10	5	50,00
19	4	2	50,00
20	11	6	45,45
21	8	5	37,50
22	9	4	55,56
23	7	5	28,57
24	6	3	50,00
25	8	2	75,00
26	11	8	27,27
27	10	6	40,00
28	8	3	62,50
29	5	4	20,00
30	12	9	25,00
31	5	3	40,00

Анализ лексического состава источника:

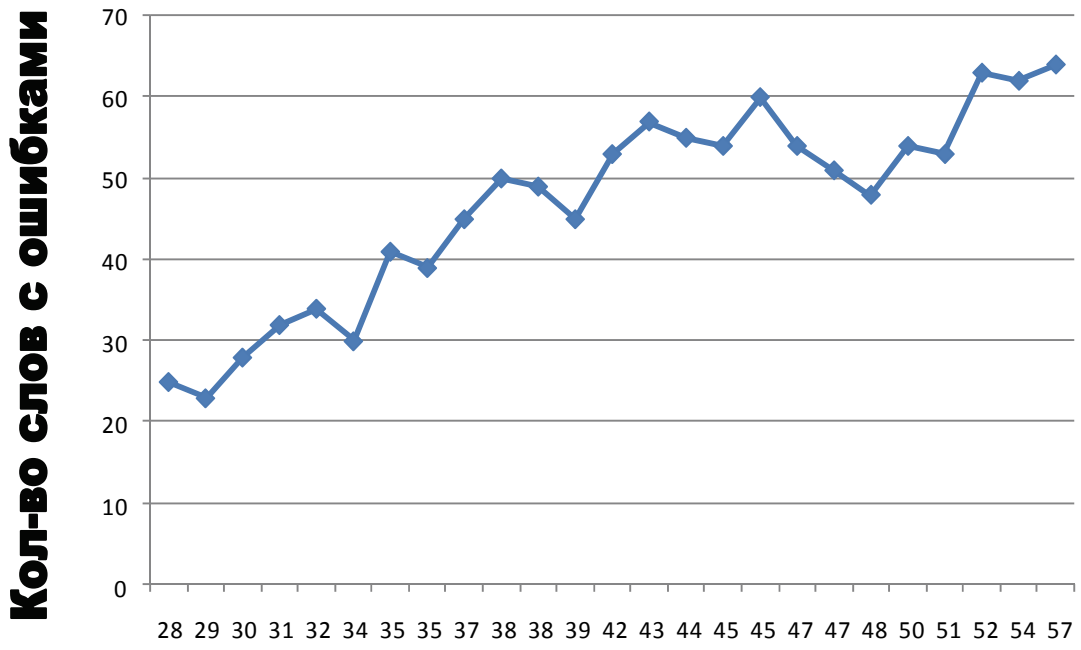
Таблица 9. Статистические данные распознавания слов

Страница	Количество слов	Количество проверяемых несловарных слов	Количество ошибочных слов	Точность распознавания слов
7	121	29	23	80,99
8	143	42	53	62,94
9	132	47	54	59,09
10	148	45	60	59,46
11	145	54	62	57,24
12	152	48	48	68,42
13	163	31	32	80,37
14	158	43	57	63,92
15	162	51	53	67,28
16	140	35	39	72,14
17	157	38	50	68,15
18	133	37	45	66,17
19	138	45	54	60,87
20	126	52	63	50,00
21	131	57	64	51,15
22	142	50	54	61,97
23	146	39	45	69,18
24	143	47	51	64,34
25	144	34	30	79,17
26	124	30	28	77,42
27	141	38	49	65,25
28	120	28	25	79,17
29	156	44	55	64,74
30	126	32	34	73,02
31	147	35	41	72,11

Из лексического анализа можно сделать следующие выводы:

- Количество несловарных слов на странице чуть более трети;
- Частота появления ошибок в несловарных словах намного выше (что вполне естественно из-за присутствия вышедших из употребления символов, таких как «Ъ», «ѣ», «ї»).

Зависимость количества ошибок от количества несловарных слов:



Из данного графика видно что при увеличении количества несловарных слов увеличивается также и количество ошибочных слов, что, как было сказано выше, является следствием присутствия символов вышедших из употребления, а также отсутствия данных слов в современном словаре.

Сводная таблица наиболее распространённых ошибок:

Таблица 10. Типы ошибок

Тип ошибки	Пример		Причина возникновения	Количество в тексте
	Ошибка	Исправление		
«т» - «ш»	рапор**ш**уют боль**т**ей	рапор**т**уют боль**ш**ей	Схожесть написания символов, большие засечки	123
«с» - «е»	ча**е**овыхъ бе**е**конечно	ча**с**овыхъ бе**с**конечно	Схожесть написания символов, не очень хорошее качество печати	100
«п» - «л»	ф**п**ангам **л**икетъ	ф**л**ангам **п**икетъ	Схожесть символов, не очень хорошее качество печати	66
«ѣ» и «ї»	в батал’он**е** т**б**мъ	в баталї**о**нѣ тѣмъ	Отсутствие символов в современном словаре, не очень хорошее качество печати	195
Лигатуры: «ми» - «лт» «ты» - «шь» «їе» - «ге»	са**лт**м пехо**шь** про**чге**	са**ми**м пехо**ты** про**чї**е	Малое расстояние между символами, не очень хорошее качество печати	76

«я» - «л» «я» - «н»	ружьл смѣяе	ружья смѣне	Схожесть символов, не очень хорошее качество печати	28

Задача 6

Провести квантитативные исследования текста: сформировать словник исследуемого источника, построить функцию распределения частот.

++

Таблица 11. Словник исследуемого источника

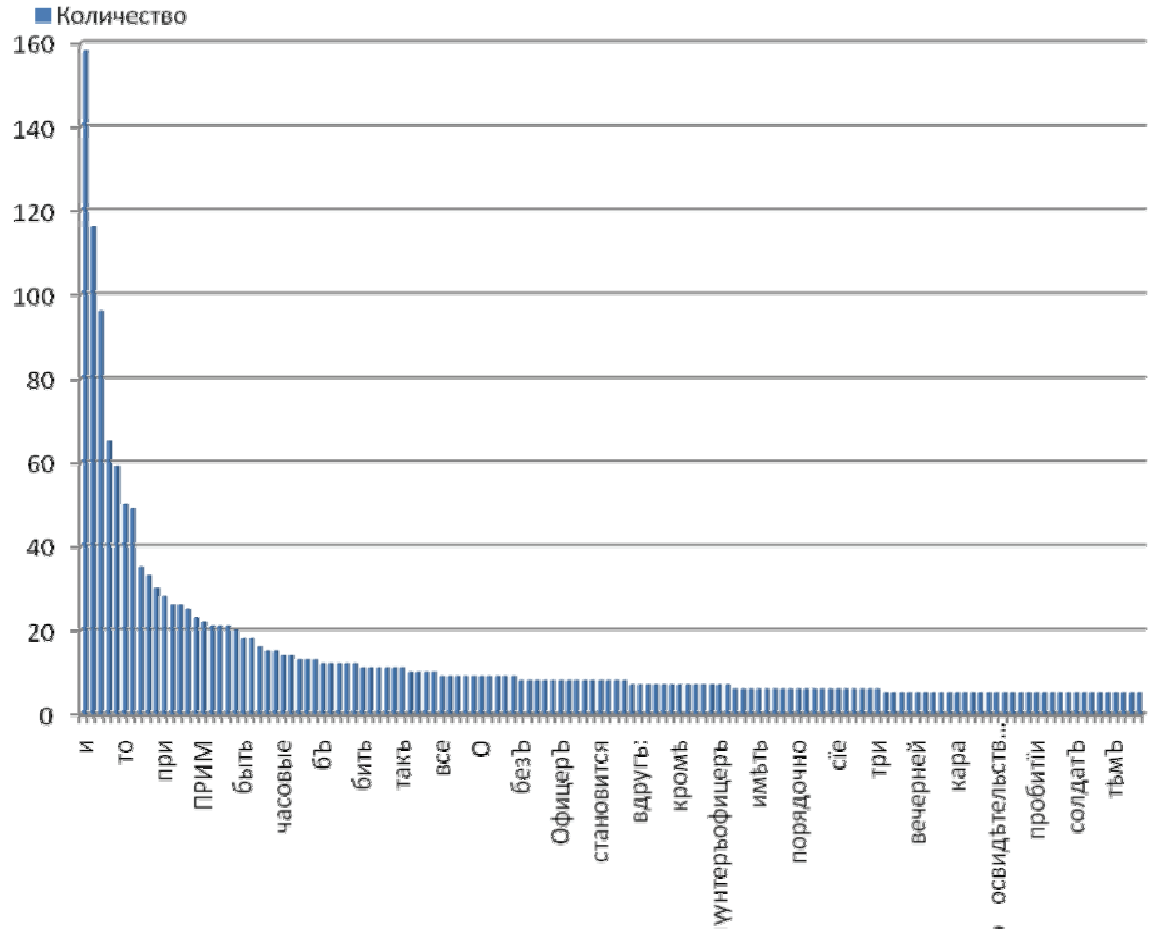
Словоформа	i	i/K	r	r^{-1}	0.1 r^{-1}
и	158	0,0433	1	1,0000	0,1000
вЪ	116	0,0318	2	0,5000	0,0500
на	96	0,0263	3	0,3333	0,0333
а	65	0,0178	4	0,2500	0,0250
по	59	0,0162	5	0,2000	0,0200
то	50	0,0137	6	0,1667	0,0167
не	49	0,0134	7	0,1429	0,0143
сЪ	35	0,0096	8	0,1250	0,0125
кЪ	33	0,0090	9	0,1111	0,0111
какЪ	30	0,0082	10	0,1000	0,0100
при	28	0,0077	11	0,0909	0,0091
у	26	0,0071	12	0,0833	0,0083
что	26	0,0071	12	0,0833	0,0083
ни	25	0,0068	13	0,0769	0,0077
или	23	0,0063	14	0,0714	0,0071
ПРИМ	22	0,0060	15	0,0667	0,0067
жЪ	21	0,0058	16	0,0625	0,0063
опЪ	21	0,0058	16	0,0625	0,0063
томЪ	21	0,0058	16	0,0625	0,0063
же	20	0,0055	17	0,0588	0,0059
быть	18	0,0049	18	0,0556	0,0056
когда	18	0,0049	18	0,0556	0,0056
полку	16	0,0044	19	0,0526	0,0053
во	15	0,0041	20	0,0500	0,0050
караулЪ	15	0,0041	20	0,0500	0,0050
часовые	14	0,0038	21	0,0476	0,0048
часовыхЪ	14	0,0038	21	0,0476	0,0048
Всѣ	13	0,0036	22	0,0455	0,0045
подЪ	13	0,0036	22	0,0455	0,0045
того	13	0,0036	22	0,0455	0,0045
бЪ	12	0,0033	23	0,0435	0,0043

всѣхЪ	12	0,0033	23	0,0435	0,0043
для	12	0,0033	23	0,0435	0,0043
караулы	12	0,0033	23	0,0435	0,0043
которой	12	0,0033	23	0,0435	0,0043
бить	11	0,0030	24	0,0417	0,0042
время	11	0,0030	24	0,0417	0,0042
за	11	0,0030	24	0,0417	0,0042
изЪ	11	0,0030	24	0,0417	0,0042
одинЪ	11	0,0030	24	0,0417	0,0042
такЪ	11	0,0030	24	0,0417	0,0042
должны	10	0,0027	25	0,0400	0,0040
ежели	10	0,0027	25	0,0400	0,0040
офицеры	10	0,0027	25	0,0400	0,0040
также	10	0,0027	25	0,0400	0,0040
все	9	0,0025	26	0,0385	0,0038
всегда	9	0,0025	26	0,0385	0,0038
выше	9	0,0025	26	0,0385	0,0038
гдѣ	9	0,0025	26	0,0385	0,0038
зари	9	0,0025	26	0,0385	0,0038
О	9	0,0025	26	0,0385	0,0038
онаго	9	0,0025	26	0,0385	0,0038
Офицеру	9	0,0025	26	0,0385	0,0038
ружье	9	0,0025	26	0,0385	0,0038
флангѣ	9	0,0025	26	0,0385	0,0038
безЪ	8	0,0022	27	0,0370	0,0037
кого	8	0,0022	27	0,0370	0,0037
ли	8	0,0022	27	0,0370	0,0037
мѣста	8	0,0022	27	0,0370	0,0037
мѣсто	8	0,0022	27	0,0370	0,0037
ОфицерЪ	8	0,0022	27	0,0370	0,0037
передЪ	8	0,0022	27	0,0370	0,0037
плечо	8	0,0022	27	0,0370	0,0037
посреди	8	0,0022	27	0,0370	0,0037
ружья	8	0,0022	27	0,0370	0,0037
становится	8	0,0022	27	0,0370	0,0037
чего	8	0,0022	27	0,0370	0,0037
чему	8	0,0022	27	0,0370	0,0037
шеренги	8	0,0022	27	0,0370	0,0037
будутЪ	7	0,0019	28	0,0357	0,0036
вдругЪ:	7	0,0019	28	0,0357	0,0036
зборЪ	7	0,0019	28	0,0357	0,0036
имѣютЪ	7	0,0019	28	0,0357	0,0036
караулахЪ	7	0,0019	28	0,0357	0,0036
командуютЪ	7	0,0019	28	0,0357	0,0036
кромѣ	7	0,0019	28	0,0357	0,0036
ПикетЪ	7	0,0019	28	0,0357	0,0036

полкахЪ	7	0,0019	28	0,0357	0,0036
правомЪ	7	0,0019	28	0,0357	0,0036
ударятЪ	7	0,0019	28	0,0357	0,0036
унтерЪофицерЪ	7	0,0019	28	0,0357	0,0036
флангамЪ	7	0,0019	28	0,0357	0,0036
бы	6	0,0016	29	0,0345	0,0034
всѣмЪ	6	0,0016	29	0,0345	0,0034
имѣетЪ	6	0,0016	29	0,0345	0,0034
имѣть	6	0,0016	29	0,0345	0,0034
караулу	6	0,0016	29	0,0345	0,0034
лагеря	6	0,0016	29	0,0345	0,0034
линїи	6	0,0016	29	0,0345	0,0034
оной	6	0,0016	29	0,0345	0,0034
порядочно	6	0,0016	29	0,0345	0,0034
право	6	0,0016	29	0,0345	0,0034
правую	6	0,0016	29	0,0345	0,0034
своему	6	0,0016	29	0,0345	0,0034
свои	6	0,0016	29	0,0345	0,0034
сїе	6	0,0016	29	0,0345	0,0034
скажетЪ	6	0,0016	29	0,0345	0,0034
становятся	6	0,0016	29	0,0345	0,0034
стоятЪ	6	0,0016	29	0,0345	0,0034
только	6	0,0016	29	0,0345	0,0034
три	6	0,0016	29	0,0345	0,0034
Барабанщики	5	0,0014	30	0,0333	0,0033
барабанщикЪ	5	0,0014	30	0,0333	0,0033
были	5	0,0014	30	0,0333	0,0033
весьма	5	0,0014	30	0,0333	0,0033
вечерней	5	0,0014	30	0,0333	0,0033
Генералитета	5	0,0014	30	0,0333	0,0033
днемЪ	5	0,0014	30	0,0333	0,0033
до	5	0,0014	30	0,0333	0,0033
долженЪ	5	0,0014	30	0,0333	0,0033
кара	5	0,0014	30	0,0333	0,0033
караулѣ	5	0,0014	30	0,0333	0,0033
кто	5	0,0014	30	0,0333	0,0033
куда	5	0,0014	30	0,0333	0,0033
опять	5	0,0014	30	0,0333	0,0033
освидѣтельствовать	5	0,0014	30	0,0333	0,0033
особливо	5	0,0014	30	0,0333	0,0033
Офицера	5	0,0014	30	0,0333	0,0033
пикеты	5	0,0014	30	0,0333	0,0033
про	5	0,0014	30	0,0333	0,0033
пробитїи	5	0,0014	30	0,0333	0,0033
роты	5	0,0014	30	0,0333	0,0033
самому	5	0,0014	30	0,0333	0,0033

сказано	5	0,0014	30	0,0333	0,0033
смѣну	5	0,0014	30	0,0333	0,0033
солдатЪ	5	0,0014	30	0,0333	0,0033
спавить	5	0,0014	30	0,0333	0,0033
споящей	5	0,0014	30	0,0333	0,0033
споящїе	5	0,0014	30	0,0333	0,0033
ступай	5	0,0014	30	0,0333	0,0033
тѣмЪ	5	0,0014	30	0,0333	0,0033
чести	5	0,0014	30	0,0333	0,0033
шагахЪ	5	0,0014	30	0,0333	0,0033
шаговЪ	5	0,0014	30	0,0333	0,0033

Функция распределения частот:



3. Выводы

В результате проведённых исследований был получен полностью отредактированный фрагмент источника «Главы к уставу о полевой службе», 1792г. Эффективность распознавания с помощью OCR-системы ABBYY FineReader 9.0 Professional Edition достаточно высока. А именно: средняя точность распознавания без дополнительных настроек равна 82,42%, а с обученным пользовательским эталоном и подключенным Словарём Академии Российской 86,17%. То есть, наилучший результат будет получен при распознавании с хорошо обученным эталоном и подключенными словарями соответствующего исторического периода источника.

По ходу работы был выявлен ряд периодически возникающих ошибок распознавания, основными причинами которых являются начертания шрифта с большими засечками, схожесть некоторых символов, невысокое качество печати 18в. и отсутствие символов в современном алфавите.

Временные затраты на непосредственно само распознавание малы по сравнению с другими видами работ (обучение эталона, редактура). Следовательно, имея готовые эталоны и словари, распознавание больших объемов схожих источников не займет длительного времени.

4. Литература

1. Филиппович Анна. Методические указания к выполнению курсовой работы по теме «Информационные технологии сохранения исторических и культурных ценностей России».
2. 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
3. ABBYY® FineReader 9 Руководство пользователя © 2008

Левкин ИУ5-95

Приложение

1. Исходная страница источни-

Генералъ-Маршемъ спускать въ полки. Стоять же онымъ у палатки по правую сторону, или какъ мѣсто дастъ.

7.

Часовыхъ же ставить въ главной квартирѣ у воротъ и дверей и гдѣ потребно будетъ: у Генерала Полнаго двухъ часовыхъ у дверей, у прочаго жъ Генералитета по одному, и ружье часовымъ держать у ноги; у Стабъ-Офицеровъ и при екипажахъ Генеральскихъ имѣть ружье на плечѣ.

ГЛАВА ВТОРАЯ,

Въ какое время смѣну чинить.

1.

Въ исходѣ девятаго часа всѣ барабанщики собираются, каждые у своихъ баталіоновъ подъ знаменами, за которыми они и барабаны свои кладутъ. И какъ скоро девять часовъ пробьетъ, то у стоящаго на правомъ флангѣ баталіона въ одинъ барабанъ ударить повѣстку, которой и во всѣхъ баталіонахъ повторяется до трехъ разъ. Послѣ чего во

ка.

2. Распознанная страница источника.

7

ГснералЪ-МаршемЪ спускать вЪ полки. Стоять же онымЪ у палатки по правую сторону , или какЪ мвсто дастЬ.

7-

ЧаеовыхЪ же ставить вЪ главной жвартирѣ у воротъ и дверей и гдѣ по-требно будетЪ : у Генерала Полнаго двухЪ часовыхЪ у дверей, упрочаго жЪ Генералитета по одному, и ружье часо-вымЪ держать у ноги-, у СтабЪ-Офице-ровЪ и при екипажахЪ ГенеральскихЪ имѣть ружье на плечѣ.

Г Л А В А В Т О Р А Я ,

Ве какое время смяіну чинить. 1.

ВЪ исходѣ девятаго часа всѣ барабанщики собираются, каждые у своихЪ баталіоновЪ подЪ знаменами, за которыми они и барабаны свои кладутЪ. И какЪ скоро девять часовЪ пробьетЪ , то у стоящаго на правомЪ флангѣ баталЇона вЪ одинЬ барабанъ ударить повестку, которой и во всѣхЬ баталЇонахЪ повторяется 40 трехЬ разЪ, Послѣ чего во

3. Отредактированная страница источника.

ГенералЪ-МаршемЪ спускать вЪ полки. Стоять же онымЪ у палатки по правую сторону, или какЪ мѣсто дастЪ.

7.

ЧасовыхЪ же ставить вЪ главной кварпирѣ у воротЪ и дверей и гдѣ потребно будетЪ : у Генерала Полнаго двухЪ часовыхЪ у дверей, у прочаго жЪ Генералитета по одному, и ружье часовымЪ держать у ноги; у СтабЪ-ОфицеровЪ и при екипажахЪ ГенеральскихЪ имѣть ружье на плечѣ.

ГЛАВА ВТОРАЯ,

ВЪ какое время смѢну чинить.

1.

ВЪ исходѣ девятаго часа всѣ барабан-щики собираются, каждые у своихЪ ба-талїоновЪ подЪ знаменами, за которы- ми они и барабаны свои кладутЪ. И какЪ скоро девять часовЪ пробьетЪ , то у стоящаго на правомЪ флангѣ баталіона вЪ одинЪ барабанЪ ударить повѣстку, которой и во всѣхЪ баталїонахЪ повто-ряется до трехЪ разЪ. Послѣ чего во