Некоторые особенности создания автоматизированной системы дешифровки исторических стенограмм ♣

А.А. Рогов, А.В. Скабин, А.Н. Талбонен, И.А. Штеркель

Петрозаводский государственный университет
rogov@psu.karelia.ru, artb00g@gmail.com, antal@sampo.ru, shterkel_ivan@psu.karelia.ru

Аннотация

В настоящее время в архивах России имеется большой объем нерасшифрованных стенографических документов. Причина – невозможность дешифровки исторических документов современными стенографистами. В течение XIX и начала XX веков стенография в России находилась в процессе становления, поэтому существующие документы зашифрованы в разных системах, к тому же современная стенография существенно отличается от исторических систем стенографии XIX века.

Описываемая информационная система призвана решить задачу описания и дешифровки исторических стенограмм, а также ввести в научных оборот новые документы. Отличительные свойства системы: учет особенностей исторической орфографии XIX и начала XX веков, учет индивидуальных знаков разных стенографистов, возможность критического анализа, использование словаря для подсказки при дешифровке текста и т.д. Разрабатываемая информационная система будет находиться в открытом доступе и предлагаться к использованию работникам архивов, научным сотрудникам, исследователям текстологам.

1. Современное состояние науки в данной области

Распознавание рукописных исторических документов стало в последние годы одним их бурно развивающихся научных направлений. Алгоритмы и программы автоматического распознавания текста разрабатываются уже несколько десятилетий. Общепризнанно, что распознавание текста включает в себя этапы предобработки (бинаризации изображения), сегментации (выделения текстовых областей, строк, слов, символов), анализа бинарных изображений символов или слов (установления значений признаков, сравнения с эталонами), и выбора наиболее подходящих словоформ из словаря в соответствии с определенной моделью языка. Можно сказать, что задача распознавания текстов на европейских языках, напечатанных на лазерных принтерах с использованием наиболее употребительных шрифтов и отсканированных на планшетных сканерах, практически решена. Однако все не так просто даже для книг середины XX века – изображения могут требовать существенной предобработки, выходящей за рамки функций, реализованных в популярных OCR-программах, шрифт может оказаться необычным, а язык - устаревшими с позиции современного словаря. Дополнительные сложности создают искривления строк, перепады яркости, просвечивания текста с обратной стороны и другие дефекты оригинала и изображения. Показательно, что один из самых заметных отечественных проектов по оцифровке печатных книг XVIII века, хранящихся в библиотеках Казани, потребовал решения комплекса проблем по устранению дефектов, сегментации, разработки специального драйвера клавиатуры, созданию грамматических правил и словаря русского языка того времени.

Сложности многократно увеличиваются при попытке решения задачи распознавания текста исторической стенограммы. Сложность данной задачи связана с рядом особенностей. Во-первых, нет людей, владеющих стенографическими записями, которыми пользовалась в XIX – начале XX вв, известны только учебники. Зачастую, стенографист может использовать свои нестандартные обозначения, так как обычно он расшифровывает записи сам. В стенографической записи применяется метод пропуска гласных букв, объединяются в один значок наиболее часто встречающиеся сочетания букв или наиболее распространенные слоги. Также, некоторые значки стенографической записи очень похожи, но могут иметь разное значение, существенное значение имеет размер символа и т. д.

В настоящее время, активно разрабатываются методы устранения дефектов и улучшения качества цифровых изображений рукописей, а также сегментации строк. Поскольку сегментация символов в рукописных текстах часто оказывается

Труды XIV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2011), Санкт-Петербург, Россия, 2011.

затруднительной, предлагаются специальные алгоритмы распознавания слитных слов и целых строк, основанные на скрытых марковских моделях и случайных полях. Большое внимание уделяется распознаванию древнегреческих текстов и арабских рукописей. Анализ публикаций и практических задач в области распознавания исторических документов позволяет сделать следующие выводы [1]:

- не существует алгоритмов распознавания, которые могли бы одинаково успешно применяться к рукописям разных языков и эпох. Все алгоритмы ориентированы на определенные типы рукописного письма;
- даже самые признанные алгоритмы не доведены до готовых программных систем распознавания рукописей, которые можно было бы использовать в отечественных проектах по оцифровке исторических документов;
- практические задачи, требующие распознавания текстов оцифрованных рукописей, в настоящее время столь часты и разнообразны, что путь разработки для каждой такой задачи специализированного алгоритмического и программного обеспечения оказывается невозможным.

В рамках исследования ставиться задача создания достаточно универсальной программной системы для автоматизированного распознавания исторических стенограмм, для которых автоматическое распознавание оказывается пока невозможным. Предлагаемая система автоматизированной дешифровки исторических стенограмм с возможностью интеллектуальной поддержки принятия решений при наборе позволит существенно ускорить процесс перевода рукописного текста в текстовый файл и повысит точность его дешифровки.

Разрабатываемая программная система будет обладать следующими особенностями:

- Ускоренный набор, по сравнению с обычным набором.
- Связь графического изображения текста и его текстового представления.
- Интеллектуализация процесса набора, которая включает поиск набираемого слова среди слов словаря и предложение возможных вариантов набора, а также анализ графического написания слова и предлагаемый вариант его дешифровки (анализ по графическому написанию фразы или целого слова, а не его отдельных букв).
- Возможность организации виртуальной клавиатуры символов, что позволит набирать текст с помощью тех символов, которые использовались при создании источника.
- Возможность автоматического распознавания в тексте отдельных часто повторяющихся фраз или слов путем поиска похожих фрагментов изображения, что существенно ускорит процесс обработки текста. Кроме того, даже выявление нескольких ключевых слов в изображении документа позволит сделать его доступным для поисковых машин.
- Возможность совместной работы по распознаванию больших коллекций текстов коллективом исследователей с единым словарем и перекрѐстной проверкой результатов.

Блок интеллектуального перевода набранного текста на современный язык будет содержать два модуля. Первый модуль – это модуль распознавания текста стенограммы, то есть перевода ее в текстовый формат в графике и орфографии оригинала. Второй модуль будет осуществлять дешифровку стенограммы на современный русский (или какой-либо другой) язык.

Модульная структура блока в случае необходимости позволит добавлять другие модули по переводу текста на современный язык. Работа данного блока должна основываться на словаре. В нашей системе по умолчанию используется база данных слов русского языка XIX века, содержащая более 50 000 словоформ. Она содержит написание слов в современной орфографии и орфографии XIX века, частоту употребления словоформ и т.д.

Предлагаемая нами автоматизированная система может быть реализована как настольное приложение. Но больший интерес, на наш взгляд, вызывает возможность ее реализации в виде Web-приложения. Это позволит, во-первых, добиться ее наибольшей универсальности и независимости от аппаратного и программного обеспечения (так как Web-браузеры есть на всех компьютерах, подключенных к Интернету), и во-вторых, организовать совместную, распределенную работу сообщества исследователей над изображениями стенограмм. Совместные исследования текстовых коллекций в рамках сетевых сообществ уже рассматривались ранее, однако использование подобных технологий на этапе расшифровки стенограмм выглядит не менее важным, так как позволяет одновременно повысить скорость распознавания и его точность за счет организации перекрѐстной проверки.

2. Описание работы автоматизированной системы

Общая блок-схема работы автоматизированной системы представлена на рисунке 1, а также на детализирующих рисунках 2 и 3.

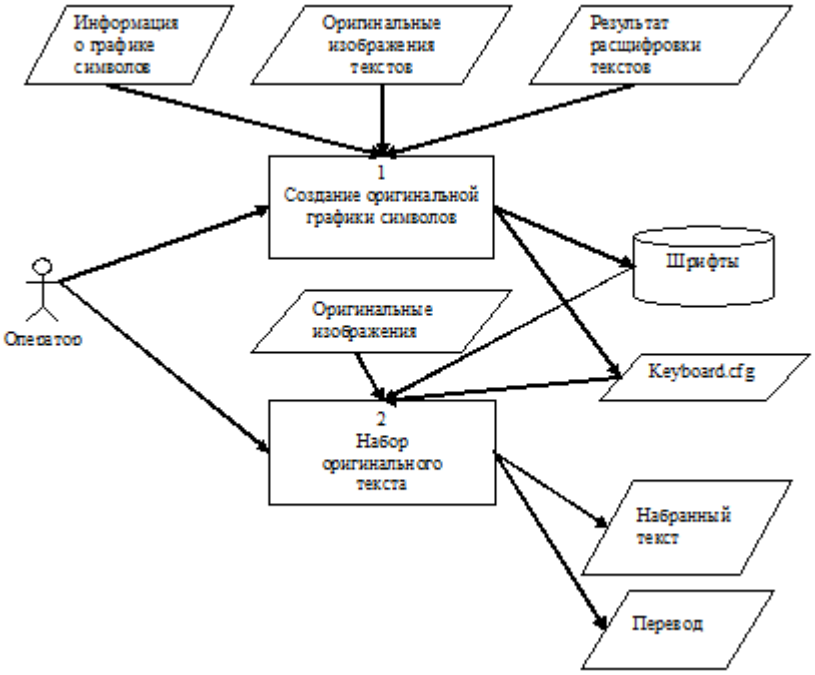

Рис. 1. Схема работы автоматизированной системы

Выделим 3 этапа работы по созданию системы и ее использованию: создание оригинальной графики символов; разработка блока дешифровки символов и набор текста.

2.1. Создание оригинальной графики символов

На данном этапе на основе информации о возможной графике символов и оригинальных изображений уже распознанных текстов, формируется виртуальная клавиатура, состоящая из набора векторных символов, встречающихся в исходном материале, сохраненных как векторные шрифты, например, в формате SVG. На рисунке 2 представлена схема процесса формирования виртуальной клавиатуры. Данный процесс состоит из последовательно выполняемых шагов представленных далее.

Предобработка. Поскольку для составления полного списка оригинальных символов требуется извлечение символов из оригинальных изображений текстов, последние необходимо подвергнуть предварительной обработке с целью устранения лишнего шума и улучшения их качества. Ввиду того, что тексты оказались очень неоднородными по яркости, было принято решение обрабатывать их частями.

Сканирование. Обработанные изображения, полученные как результат выполнения предыдущего процесса, подвергаются сканированию. Данное сканирование заключается в поиске и выделении отдельных графем на изображении так, чтобы каждому символу текста соответствовала графема. В результате выполнения данного процесса получается список первичных растров. В результате экспериментов было подтверждено утверждение в [2], что для выделения символов можно использовать пороговый метод по яркости. Порог выбирался, так, чтобы суммарное число черных пикселей составлял примерно 12% от общего числа. После этого символы разбиваются на отдельные, каждый из которых записывается в отдельный файл. Критерием разбивания служит линейная связность.

Коррекция растров. Оператор просматривает первичные растры с целью выявить какие-либо ошибки и устраняет их. В результате получается список растровых символов,

Векторизация. Растровые символы анализируются и из них формируются векторные аналоги. На выходе – список первичных шрифтов.

Коррекция шрифтов. Оператор просматривает список первичных шрифтов и исправляет обнаруженные дефекты. В результате получается список шрифтов.

Сопоставление. На основе набора шрифтов, информации о возможной графике символов и распознанных оригинальных текстов оператор создает таблицу соответствия между векторными символами и лексемами обычного текста.

Формирование виртуальной клавиатуры. На основе таблицы соответствия оператор с помощью данной системы формирует конфигурационный файл клавиатуры Keyboard.cfg.

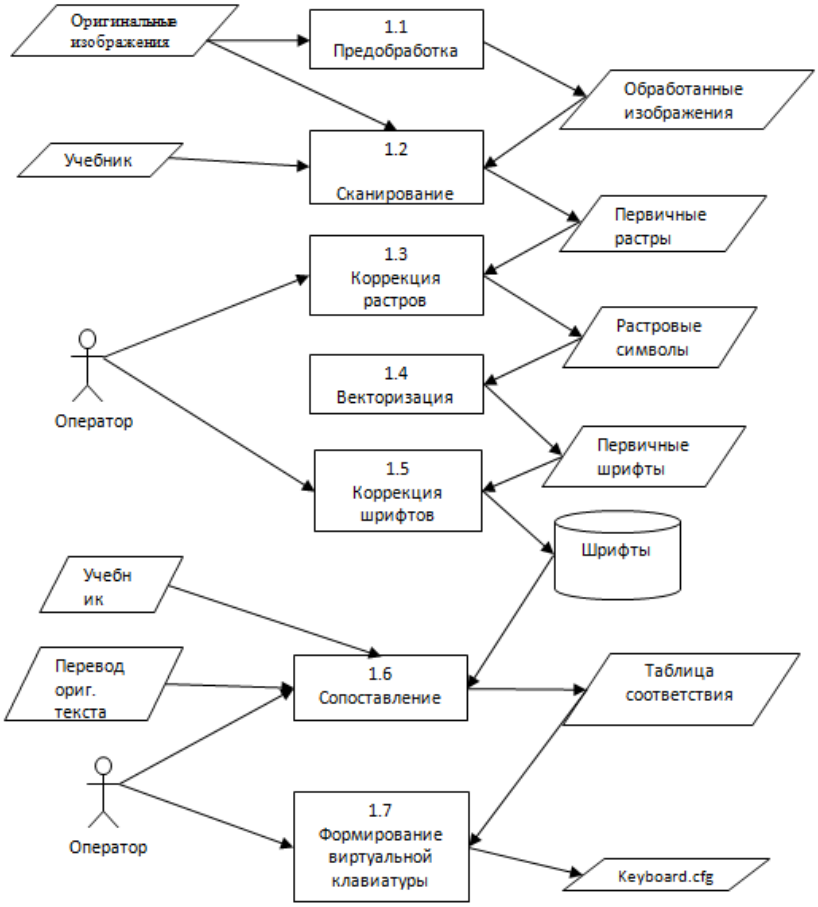

Рис. 2. Схема формирования виртуальной клавиатуры

2.2. Блок дешифровки символов

Виртуальная клавиатура содержит в себе сопоставление между оригинальными символами (графемами) и буквосочетаниями (лексемами). Данная информация записана как конфигурационный файл виртуальной клавиатуры Keyboard.cfg. С помощью данного конфигурационного файла и набора векторных шрифтов пользователю предоставляется возможность набирать текст оригинальными символами и при этом параллельно формировать перевод текста.

2.3. Набор оригинального текста.

Набор текста осуществляется оператором с помощью созданной ранее виртуальной клавиатуры. В процессе набора символов система формирует и отображает пользователю список вариантов текущего набираемого слова, а также варианты перевода данного слова. Кроме того, система проверяет набранный текст на наличие синтаксических или стилистических ошибок и сообщает пользователю информацию об ошибках. В результате выполнения данного процесса система формирует текст, набранный оригинальными символами, а также перевод данного текста. На рисунке 3 представлена схема процесса набора текста. Данный процесс состоит из последовательно выполняемых шагов.

Предобработка. Оригинальное изображение перед непосредственным набором проходит предварительную обработку с целью устранения нежелательного шума и повышения его качества.

Набор текста. Обработанное изображение подается в модуль набора текста системы, и на основании данного изображения оператор выполняет набор текста. В процессе набора система предлагает пользователю различные варианты набора и сообщает ему об ошибках. На основе этой информации пользователь решает, что делать дальше с текущим словом. После всех операций набора, выбора вариантов и корректировки на выходе формируется набранный оригинальными символами текст в виде изображения и текстовый перевод.

Интеллектуальный набор. Система предлагает пользователю возможные варианты набора на основе введенных символов. Выход данного процесса, варианты слова, подается на вход основному процессу, набору текста. Таким образом, обеспечивается обратная связь, используемая для коррекции набранного текста.

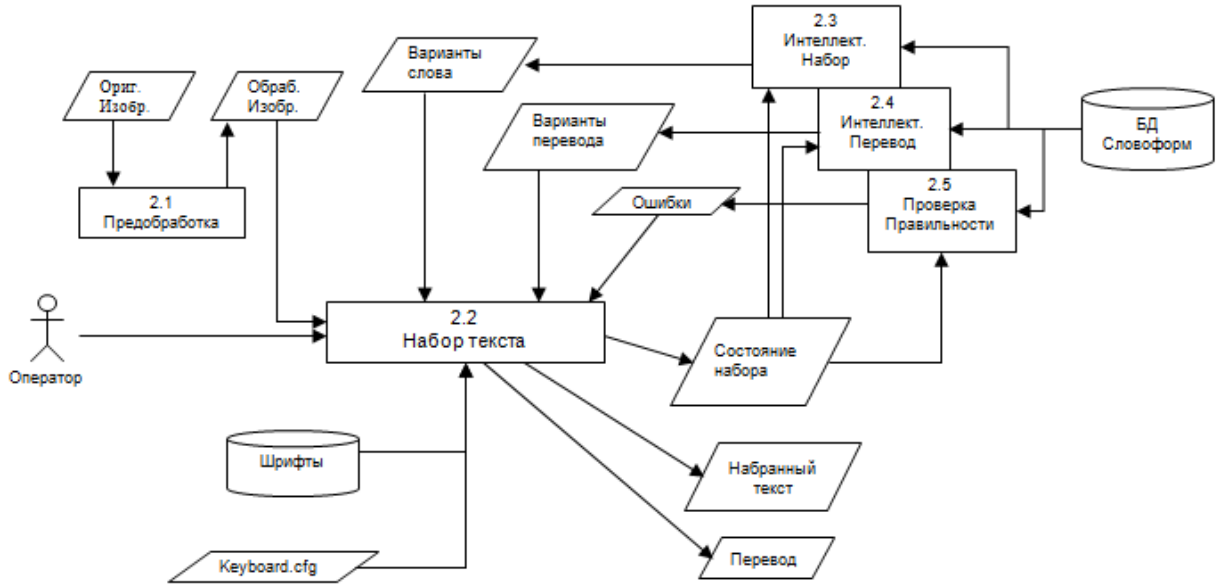

Рис.3. Схема набора текста

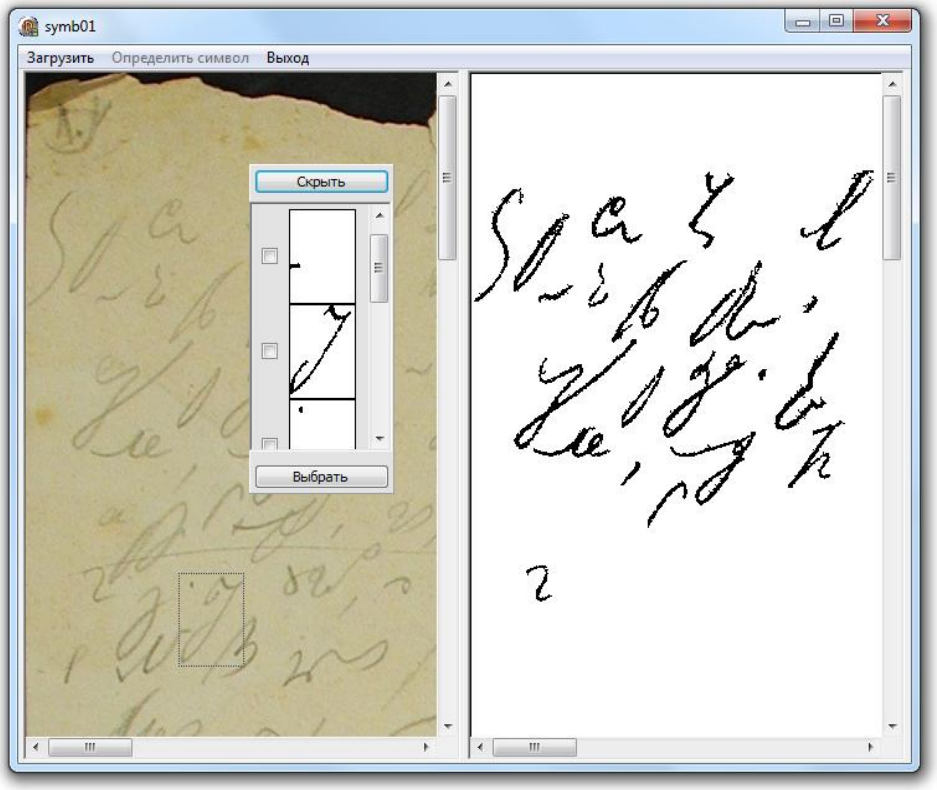

Рис. 4. Интерфейс модуля создания оригинальной графики символов

Интеллектуальный перевод. Система анализирует графическое написание целого слова (а не отдельных букв) и предлагает пользователю варианты его дешифровки (перевода).

Проверка правильности. Система проверяет синтктику и стилистику текста и сообщает пользователю о выявленных ошибках.

3. Описание интерфейса модуля создания оригинальной графики символов

Интерфейс программы можно разделить на 2 области (см. рис. 4). Первая область – это область рукописного текста, которая содержит в оригинальное изображение. На данной области оператор выделяет один символ рукописного текста, после чего ему будет предложено выбрать один из распознанных программой сегментов

рукописного текста, который будет в наибольшей степени соответствовать рукописному символу. После выбора, сегмент преобразуется в растровое изображение, и записывается в базу символов и отображается на 2-ой области модуля. Если в результате обработки выделенной оператором области получается один сегмент, то он отображается на 2-ой области модуля и записывается в базу символов. Вторая область – область, на которой отображаются растровые изображения символов, полученые в процессе обработки изображения. При загрузке оригинального изображения, во 2-ой области отображаются все символы, которые были ранее получены при обработке данного изображения. А так же в этой области отображаются символы, которые были получены при текущей обработки изображения. Данная область необходима, для того чтобы оператору было легче ориентироваться, какой символ был обработан, а какого символа еще нет в базе символов.

В дальнейшем все полученные символы будут использованы в конфигурационном файле для создания виртуальной клавиатуры.

4. Описание интерфейса модуля набора текста

Интерфейс программы можно разделить на 4 области (см. рис. 5).

Область рукописного текста, содержит оригинальное изображение.

Область растровых изображений символов, в данном поле содержаться символы, которые уже были введены оператором с виртуальной клавиатуры. Данная область для лучшего визуального восприятия оператора, чтобы оператор видел какие символы он уже ввел с виртуальной клавиатуры.

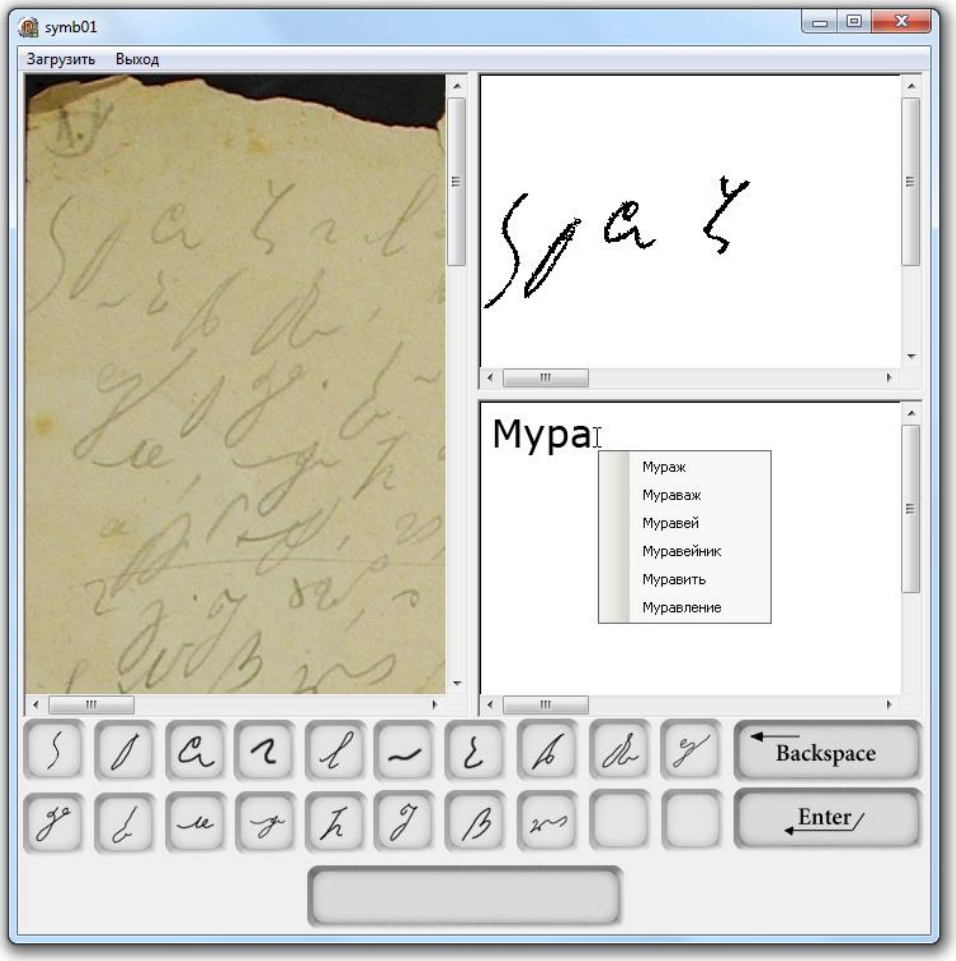

Рис. 5. Интерфейс модуля набора текста.
Пример ввода символов с выпадающим списком вариантов слов

Область печатного или распознанного текста, представляющая собой текстовое поле, содержащее соответствующие лексемы набранных символов, организованные в слова. По мере ввода символов программа автоматически вводит расшифрованные биграммы в текстовое поле на соответствующее место (как правило в конец). Для текстового поля существует свой курсор (текстовый курсор), перемещение которого строго соответствует логическим перемещениям символьного курсора.

Т.е., текстовый курсор выполняет перемещение только в том случае, когда символьный курсор перемещается между символами. Пиксельные сдвиги символьного курсора не учитываются. Каждый последующий ввод символа за исключением пробела добавляет к текущему слову соответствующую биграмму или другой набор символов, соответствующий введенному символу. Область печатного текста логически выделяет текущее слово и по запросу пользователя осуществляет поиск вариантов слов из специального словаря, соответствующих текущему слову (как правило, совпадающих, либо имеющих различие в одном символе; сравнение слов осуществляется посимвольно, начиная с начала). При выборе нужного слова, программа автоматически определяет недостающие символы и добавляет их в область отображения растровых изображений символов, а также исправляет ошибочные символы (если выбранное слово отличалось от текущего одним символом). После этого в конец рукописного текста добавляется пробел и курсор перемещается на новое место. Далее пользователь может, перемещая курсор редактировать набранные символы по отдельности, изменяя их расположение и размер, или непосредственно изменяя символы. Дальнейший ввод будет добавлять расшифрованные биграммы в новое текущее слово и так далее, пока весь текст не будет расшифрован.

Панель наборных символов, состоящая из нескольких рядов кнопок с изображенными символами ввода. В случае большого количества наборных символов и недостаточным размером панели возможна прокрутка панели к нужному месту, чтобы сделать доступным недостающий набор символов. Кроме специальных символов ввода панель содержит пробел для логического разделения слов, «ввод» для логического разделения строк и «удалить» для удаления последнего введенного символа. Ввод рукописного текста осуществляется путем комбинированного выбора символов (с помощью кликов мышкой) и позиционирования символьного курсора (с помощью специальных клавиш). Панель символов ввода формируется динамически с помощью заранее определенного набора символов (дефолтный набор). Дефолтный набор состоит из глифов символов ввода и соответствующих им символов печатного текста в виде биграмм или других буквосочетаний. Программа формирует матрицу кнопок с динамической привязкой к дефолтному набору символов, что позволяет использовать множество различных алфавитов для ввода. В процессе работы программы при каждом вводе рукописного символа, программа с помощью привязки определяет пару <рукописный символ, буквосочетание>, соответствующую нажатой кнопке и затем добавляет элементы этой пары в соответствующий текст.

В дальнейшем планируется реализация системы распознавания рукописных исторических документов в виде Web-сервиса, для организации распределенной, удаленной работы со стенограммами.

Литература

[1] Рогов, А.А. Автоматизированная система распознавания рукописных исторических документов / Рогов А.А., Талбонен А.Н., Варфоломеев А.Г. // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XII Всероссийской научной конференции RCDL’2010 (Казань, Россия, 13–17 октября 2010 г.). – Казань: Казан. ун-т РАН, 2010. - С. 469-475.

[2] Горский, Н. Распознавание рукописного текста: от теории к практике / Горский Н., Анисимов В., Горская Л.. – СПб.: Политехника, 1997.

Some features in design of the automated system of deciphering shorthand report

A.A. Rogov, A.V. Skabin, A.N. Talbonen, I.A. Shterkel

Nowadays there are many shorthand reports in Russian archives that not deciphered. The main reason why the reports are not deciphered is the absence of stenografers who can do it. The XIX and the first part of XX centuries were the period of formation of Russian stenography so many stenography systems existed at the same time. Because of this, the existing reports are ciphered in different stenography systems. Also modern stenography considerably differs from historical systems of XIX century.

The described information system is aimed to solve a problem of deciphering historical shorthand reports. Some of the distinctive system features are consideration of orthographical specifics of XIX and the first part of XX centuries, taking into account the individual signs of different stenographers, the ability of critical analysis, the possibility to use the vocabulary for assistance during the report deciphering process, etc. The developed system will be in free access for archives employees, science workers, textology researchers.

♣ Данная научная разработка поддержана грантом РГНФ № 11-01-12026в (руководитель А.А. Рогов)