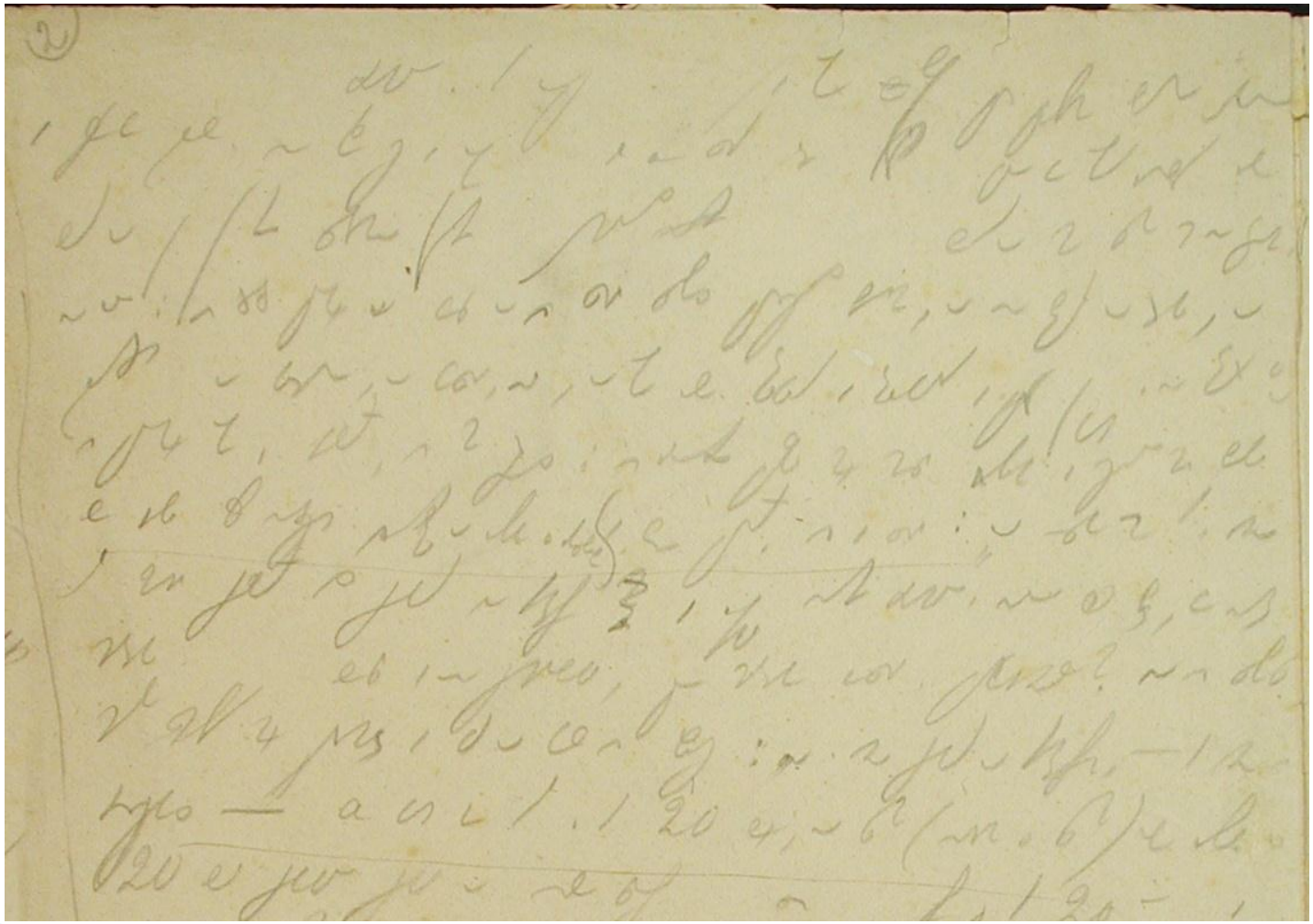
Автоматизированная система распознания рукописных исторических документов

Скабин А.В.
Штеркель И.А.

Петрозаводский Государственный Университет

- 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.

Фрагмент стенограммы



Основные этапы распознавания текста

- Предобработка;
- Сегментация;
- Анализ изображений символов или слов;
- Дешифровка содержания (выбор наиболее подходящих словоформ из словаря)

Цель работы

Создание универсальной программной системы для автоматизированного распознавания исторических рукописных текстов, включая исторические стенограммы XIX и начала XX веков.

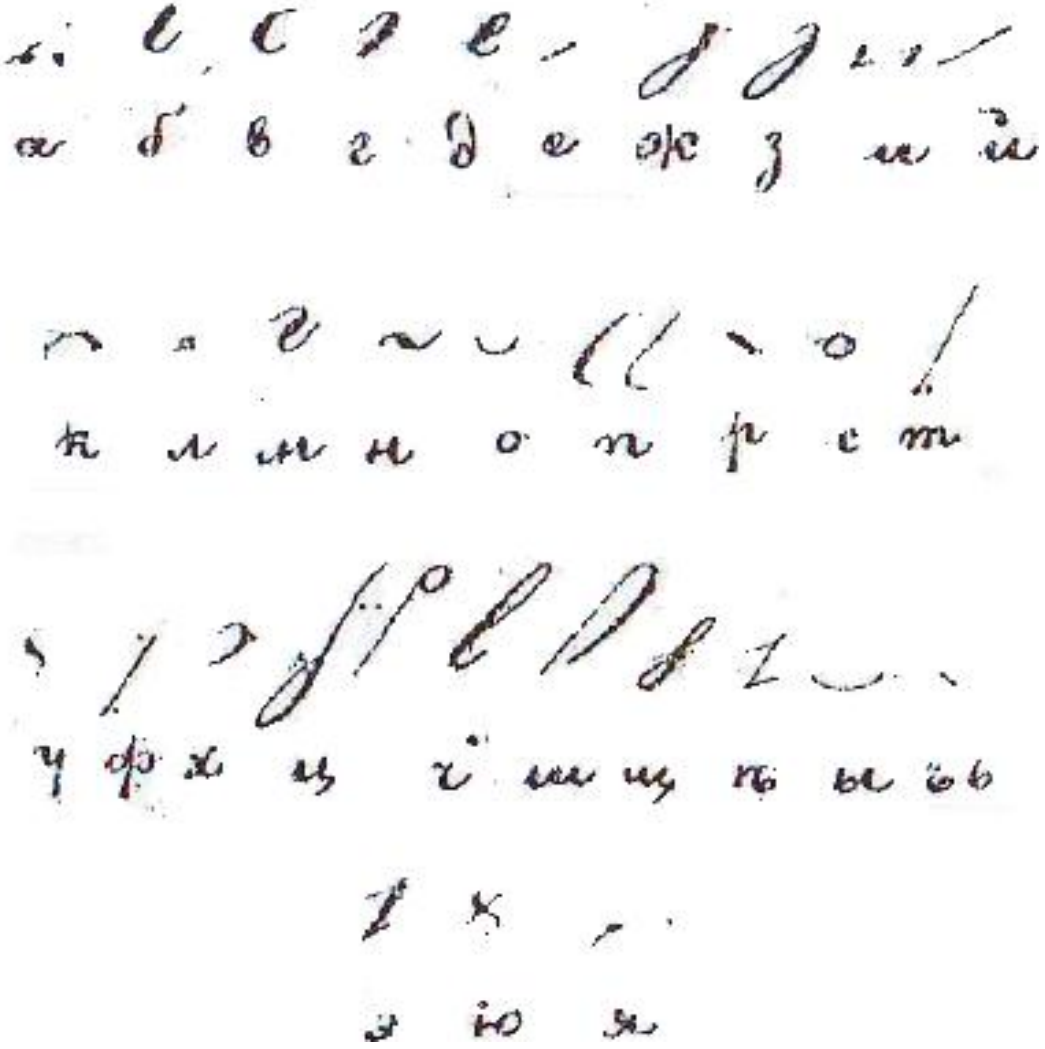
Основные сложности при дешифровке стенограмм

- Отсутствие людей, знающих особенности стенографического письма XIX и начала XX в.
- Использование нестандартных обозначений
- Использование метода пропуска гласных букв
- Наличие похожих графем с различным значением
- Наличие символов – заменителей частых символов

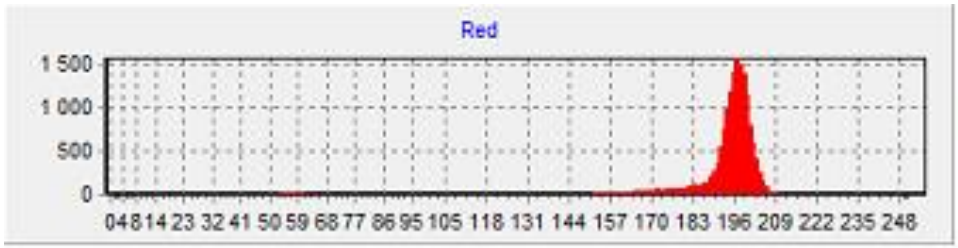
Основные характеристики системы

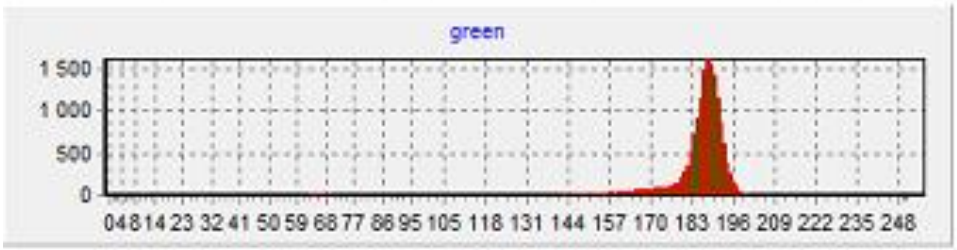
- Система автоматически контролирует состояние набора и в интерактивном режиме выдает информацию пользователю
- Система возвращает пользователю:
  варианты набора словоформы, упорядоченные по частоте встречаемости в базе данных
  варианты дешифровки словоформы
  информация об отсутствии набранного слова в базе данных

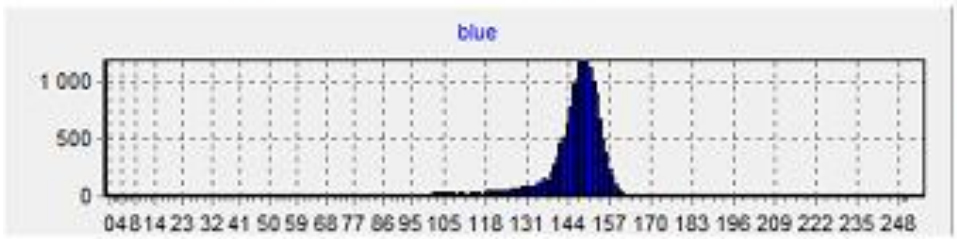
Основные знаки



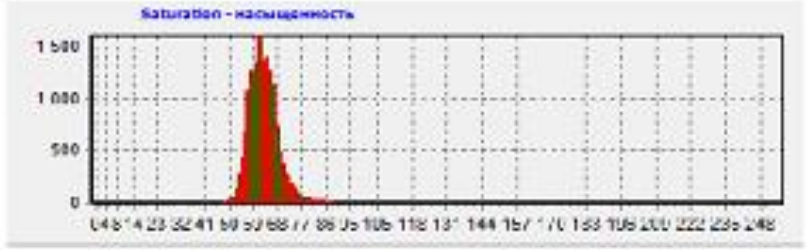
Гистограммы RGB и HSB

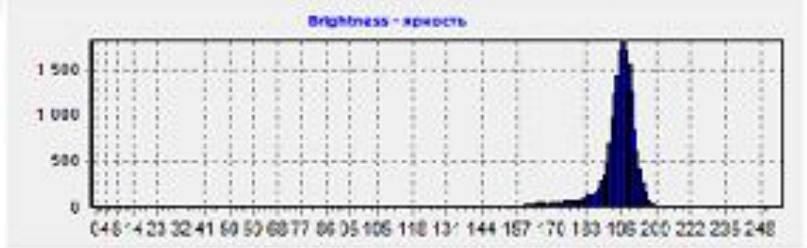






а.





б.

Бинаризация фрагмента

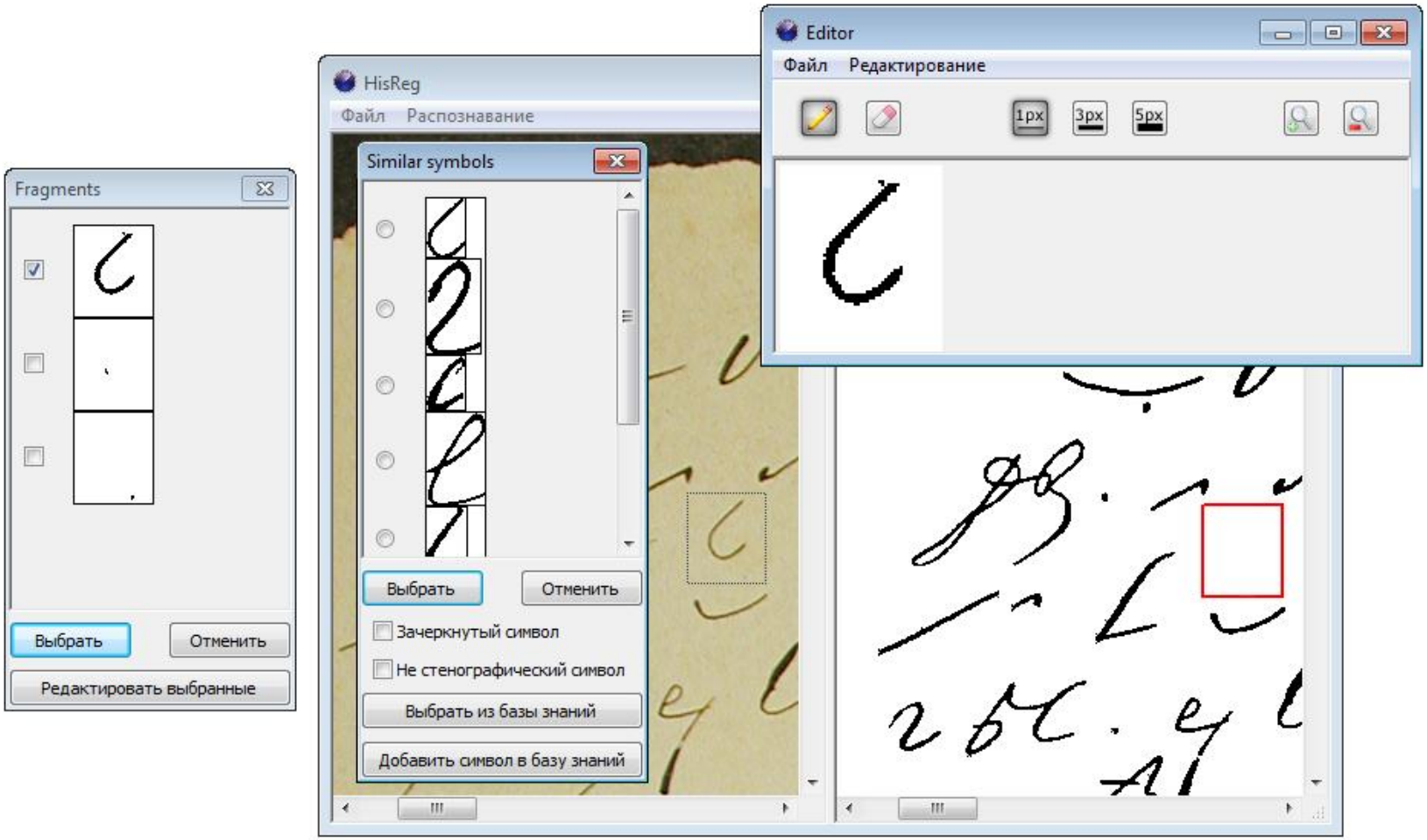
- Порог бинаризации выбирался по яркости (Brightness) так чтобы:

$$\frac{S_b}{S} \cdot 100\% \sim 13\%$$

S_b – количество черных пикселей после бинаризации

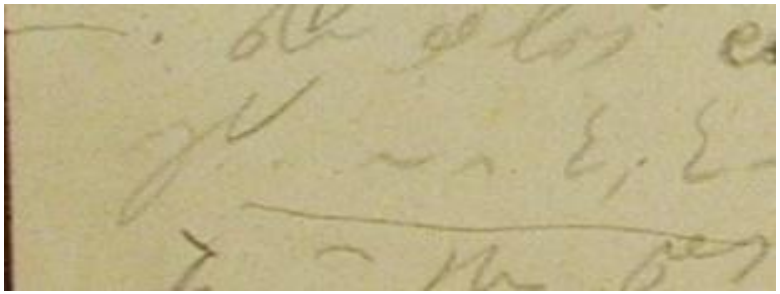
S – общее число пикселей фрагмента

Модуль создания оригинальной графики символов

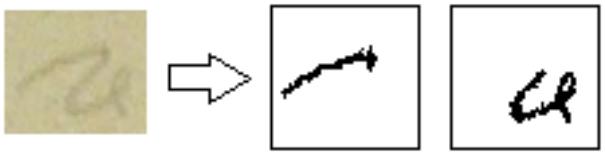


Основные сложности при создании оригинальной графики

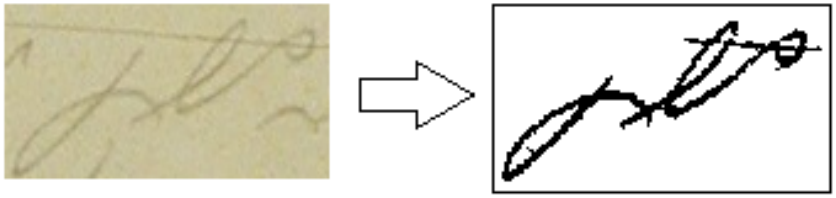
- Состаренность стенограмм;



- Разрыв символов при обработке;



- Слияние символов в слова при письме

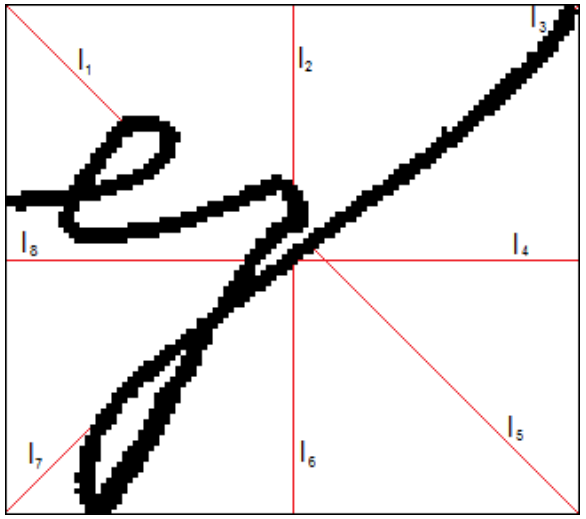


Поиск символа в базе

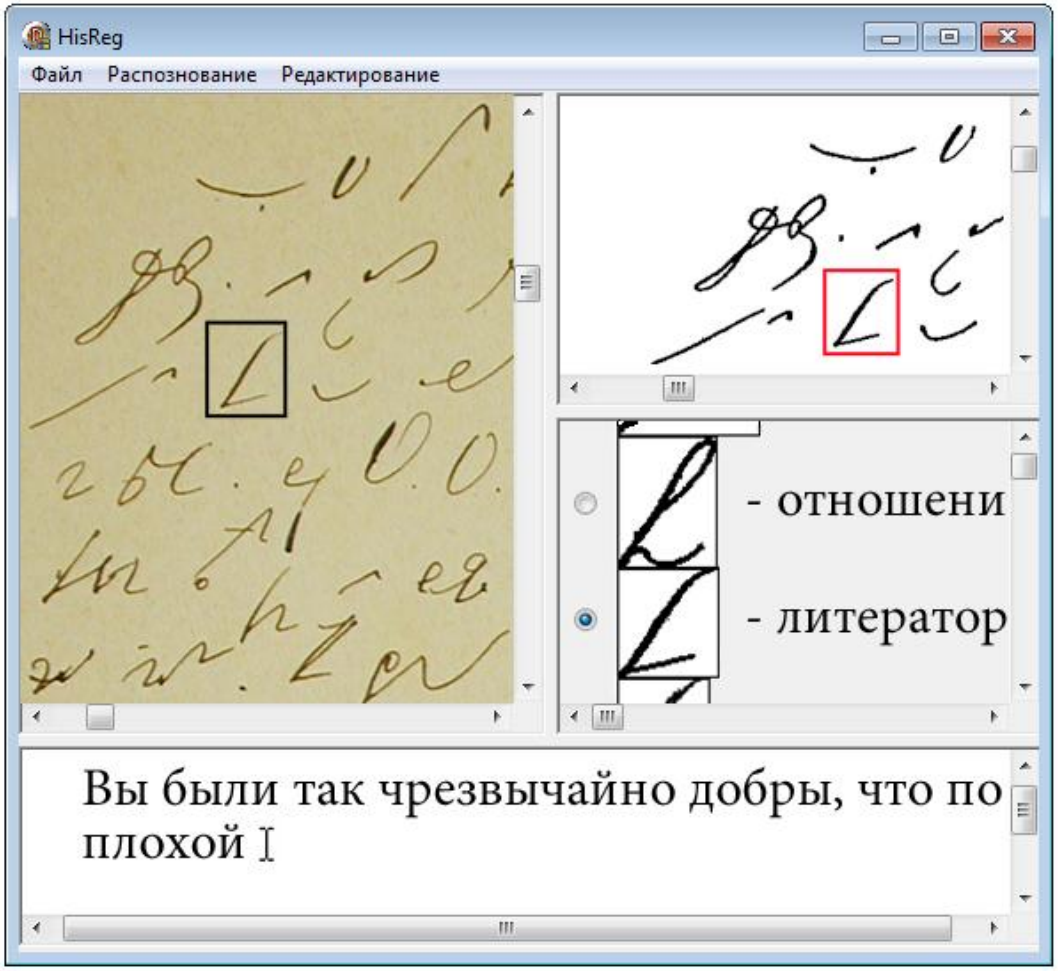
Характеристика Метод	Время сравнения	Точность
Сравнение с эталоном	3 сек. (в зависимости от размера символа)	Менее 30%
Сравнение со скелетом эталона	1-2 сек. (в зависимости от размера символа)	~40%
Метод краевых расстояний	менее 0.01 сек.	Более 60%

Метод краевых расстояний

- Измеряются расстояния $\{l_1, l_2, \ldots, l_8\}$
- Из базы знаний, выбираются такие символы у которых данные расстояния $\{\grave{l_1}, \grave{l_2}, \ldots, \grave{l_8}\}$ находятся в промежутке $(l_1 \cdot k - \varepsilon, l \cdot k + \varepsilon)$ где k – отношение высоты к ширине текущего символа, $\varepsilon = k \cdot l \cdot \alpha$, где $\alpha = 0.1$



Пример интерфейса прототипа программной системы



Основные преимущества

- Ускоренный набор
- Связь графического изображения текста и его текстового представления
- Интеллектуализированный набор
- Возможность автоматического распознавания в тексте похожих элементов
- Возможность совместной работы нескольких людей с одним словарем

Спасибо за внимание