Методика распознавания древнерусских скорописных текстов

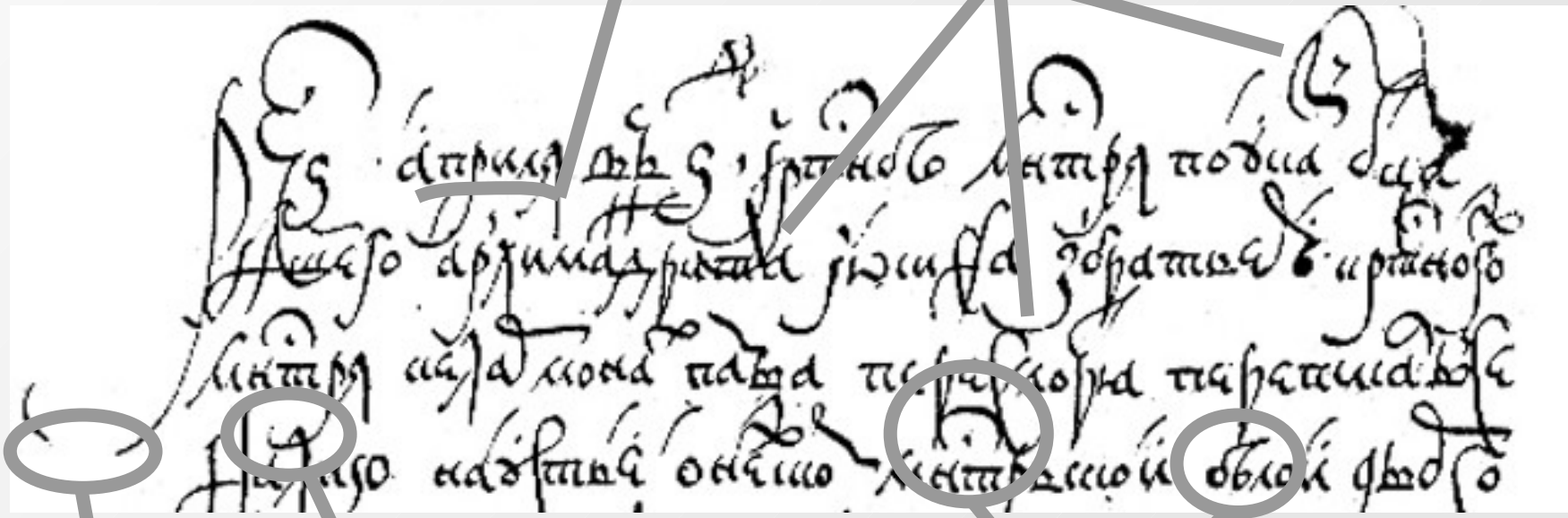
Цель:

- разработка методики автоматизированного перевода растровых изображений древнерусских скорописных текстов в электронное текстовое представление.

Актуальность:

- необходимость представления рукописей в виде электронного текста для обеспечения их компьютерной обработки;
- ограниченность круга людей, способных к чтению данных рукописей, учёными-исследователями русского языка;
- трудоёмкость ручного перевода рукописей в электронное представление;
- отсутствие современных средств распознавания, работающих с древнерусской скорописью.

Особенности скорописного способа формирования текста



Искривление линии слова

Декоративные росчерки

Дефект

Пересечение элементов букв

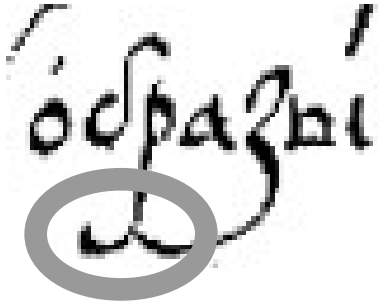
Соединение букв

Особенности скорописного способа формирования текста

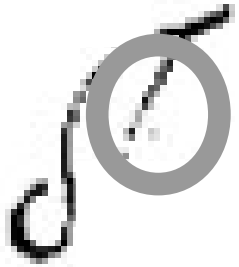
Вариативность начертания символов

Особенности скорописного способа формирования текста

Отсутствует пересечение



Лишнее пересечение

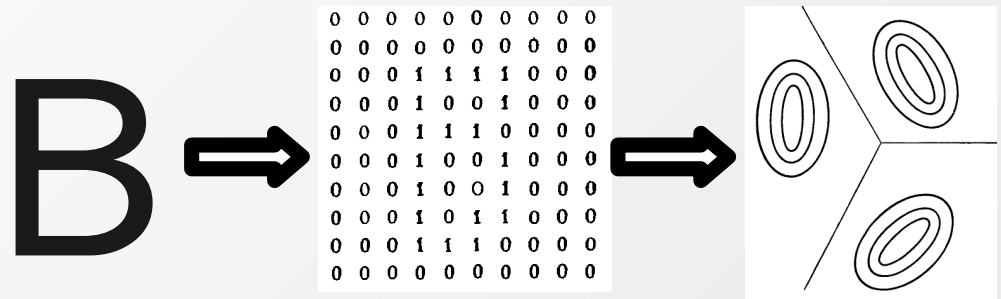


Декоративный элемент

Дефекты

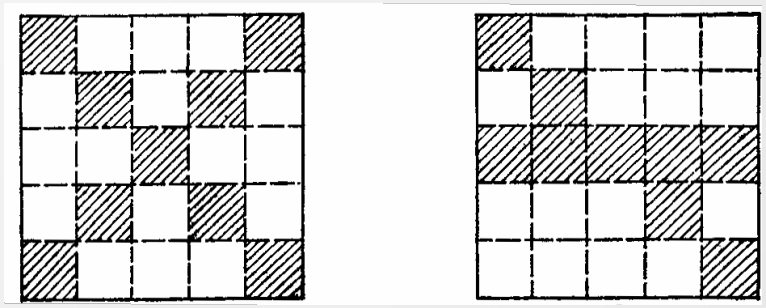
Анализ существующих методов распознавания

Евклидово пространство



- **Евклидово пространство:** представление объектов в виде наборов измерений;
- **Признаковые методы:** представление объектов в виде совокупности признаков.

Признаковые методы



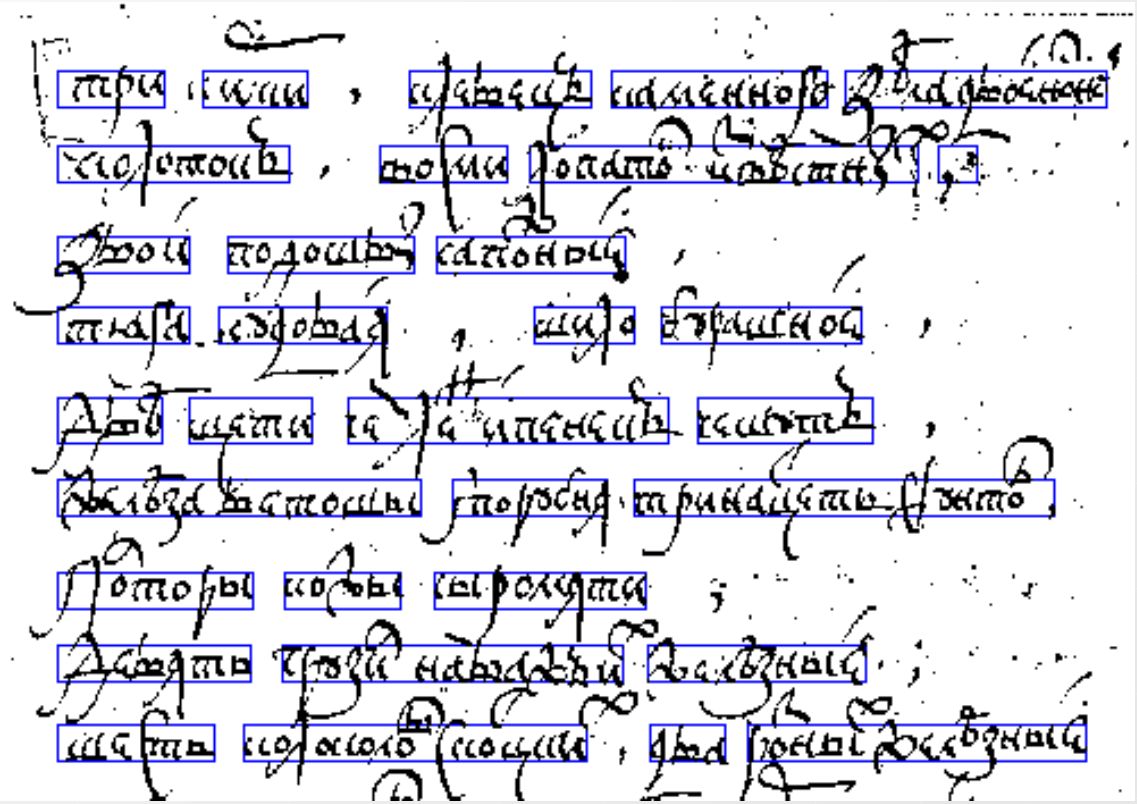
$$x_i = \{x_{i1}, x_{i2}, \ldots, x_{ik}\}$$

Основные недостатки:

- необходимость выделения объектов из общей картины для их распознавания;
- в рукописях отсутствуют явные границы символов и возможны случайные пересечения;
- применительно к рукописи выделение символа или слова равносильно его распознаванию.

Анализ существующих методов распознавания

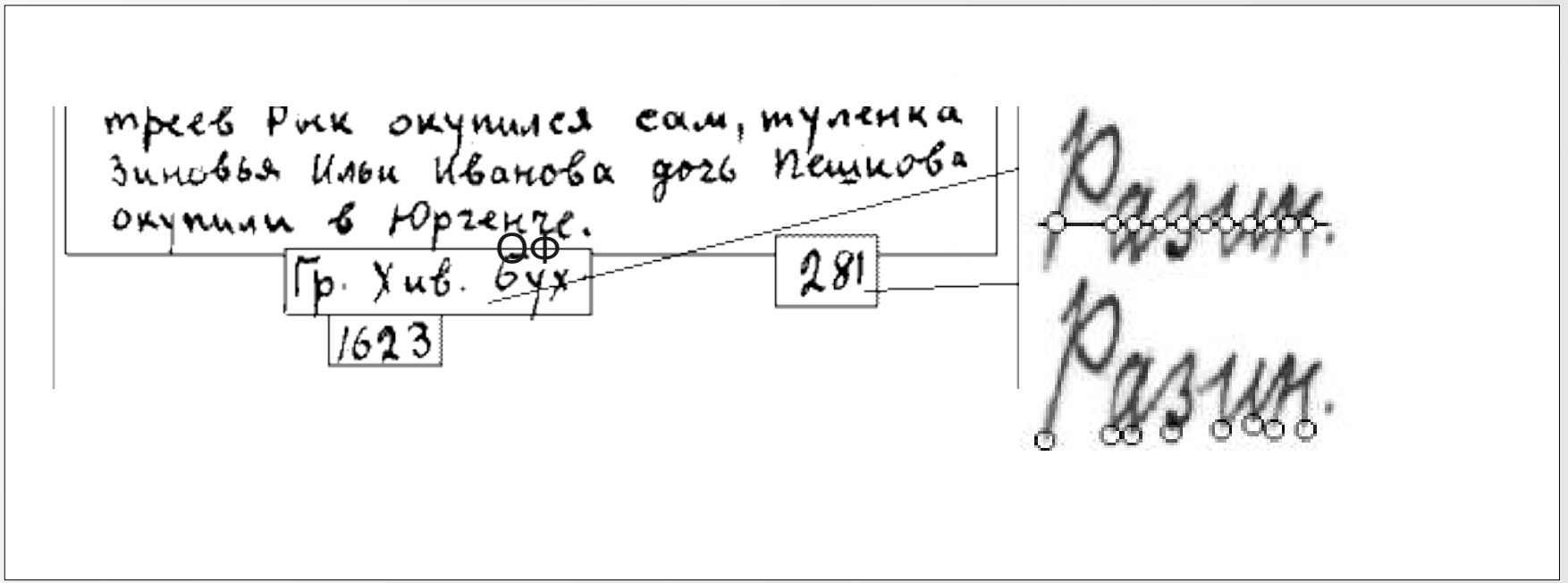
Работы Чикунова: выделение слов в тексте. Статистический подход на основе горизонтального и вертикального сканирования точек изображения



- Отсутствует выделение букв в словах;
- Выделенная область слова могут содержать не все буквы этого слова и содержать посторонние элементы.

Анализ существующих методов распознавания

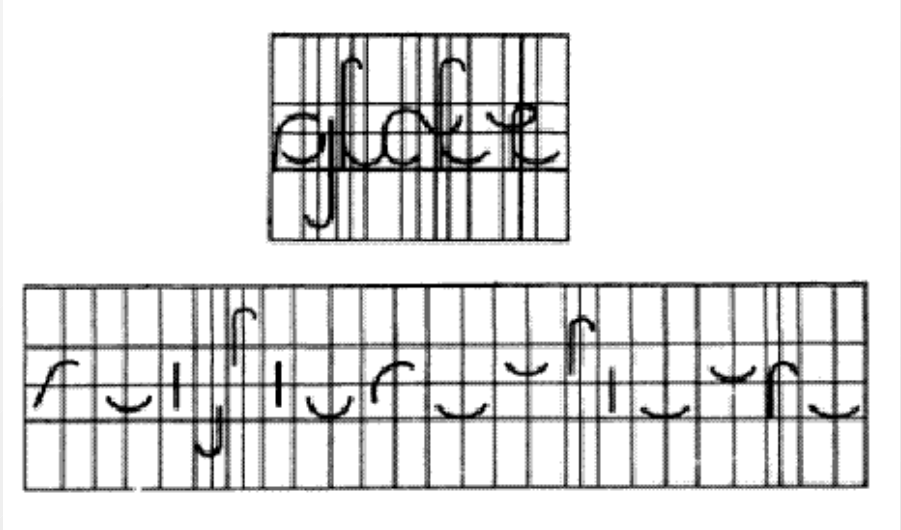
Работы Кадакина: распознавание рукописных шрифтов источников.
Признаковый подход на основе подсчёта чёрных точек
в характеристических областях слов.



- Требование к предварительному выделению распознаваемых слов.

Методы распознавания

Структурные методы

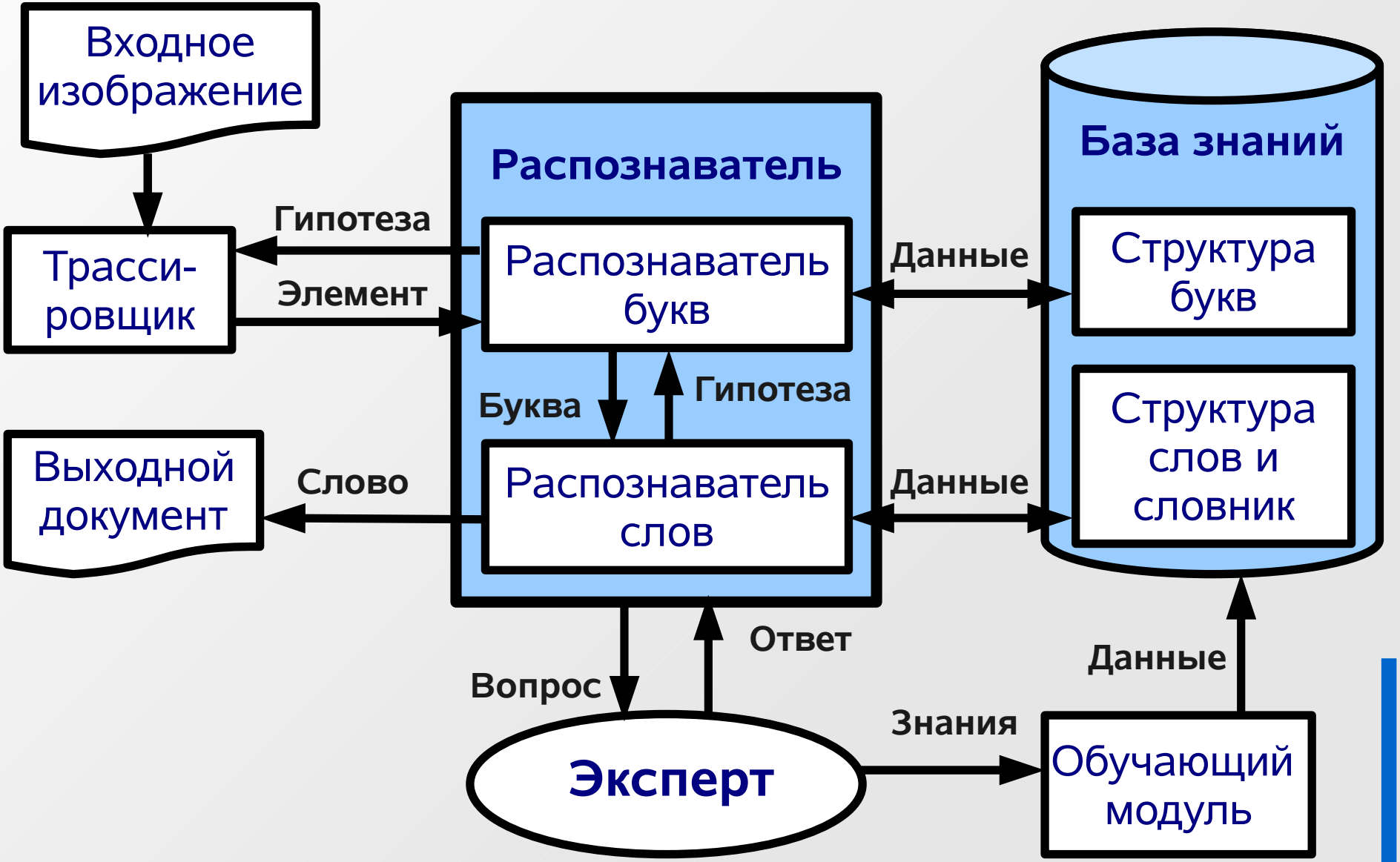


- основываются на выделении и анализе составных частей распознаваемых объектов;
- позволяют целенаправленно перемещаться по изображению в процессе распознавания;
- позволяют учитывать декоративные элементы.

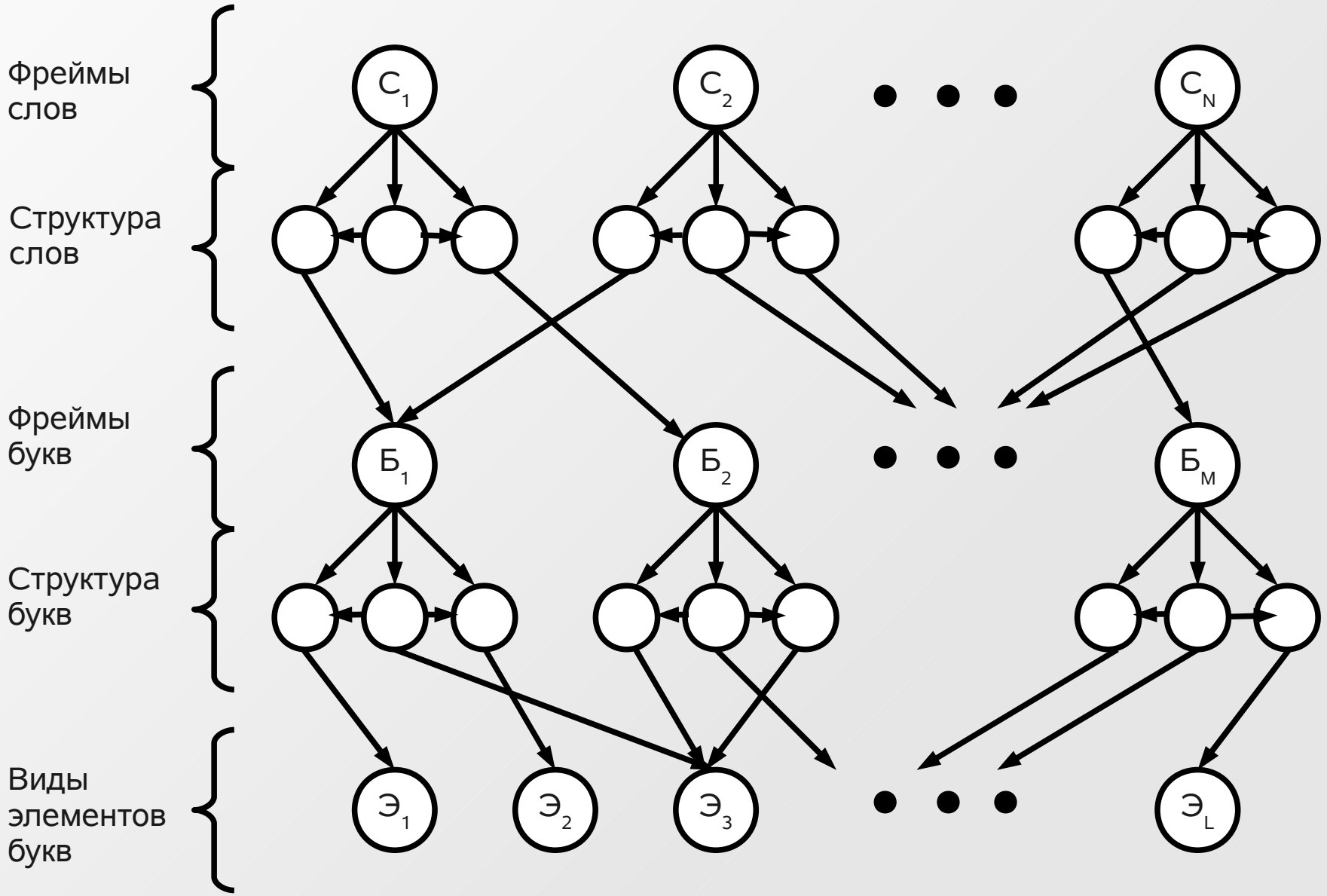
Ключевые особенности предлагаемого метода распознавания

- Структурный подход;
- Нечёткое описание структурных элементов и отношений между ними;
- Распознавание под управлением гипотез;
- Экспертный подход.

Система распознавания. Структурная схема



Структура базы знаний



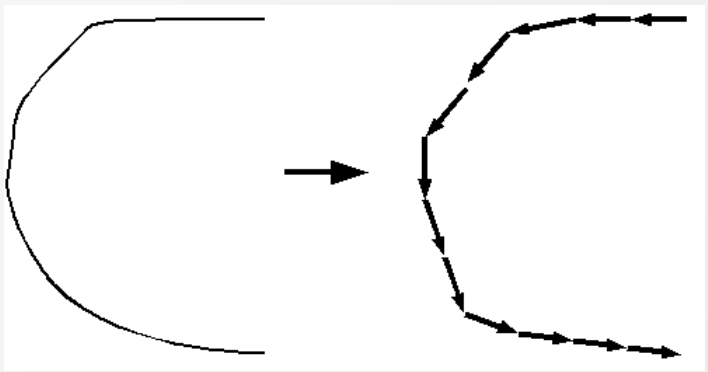
Типы узлов базы знаний

- Элементы слов:
  - Буквы;
- Структура слова:
  - Вхождения букв;
  - Пространственные отношения.

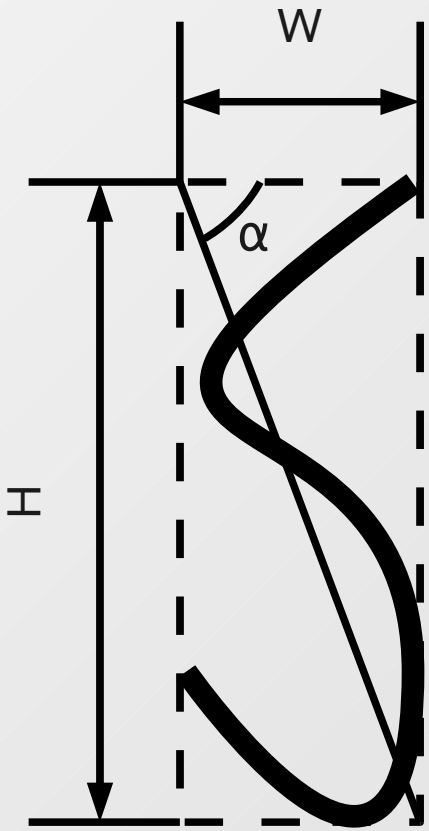
- Элементы букв:
  - Линии;
  - Точки пересечения;
- Структура буквы:
  - Вхождения линий;
  - Вхождения точек пересечений;
  - Отношения принадлежности точек;
  - Отношения соответствия точек;
  - Пространственные отношения.

Структурные элементы букв

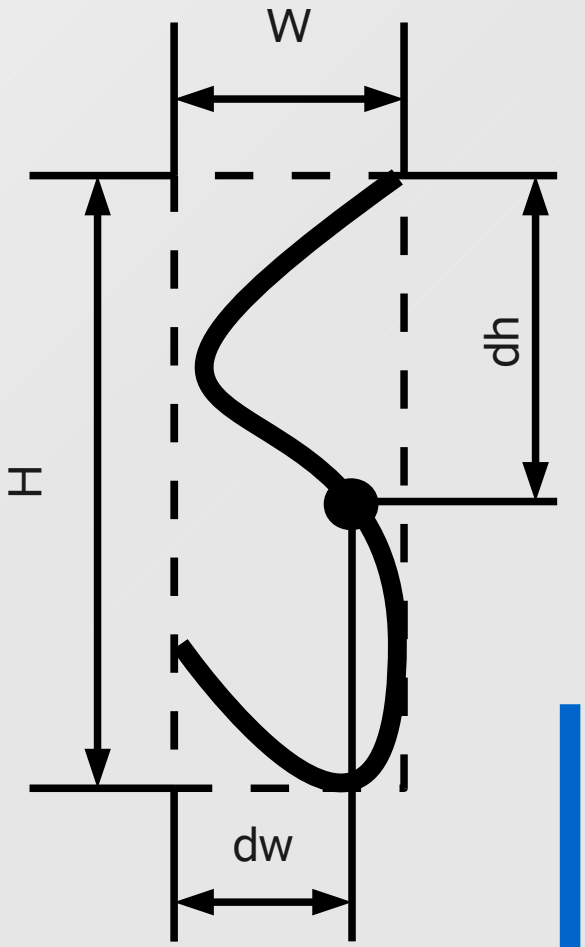
Путь линии



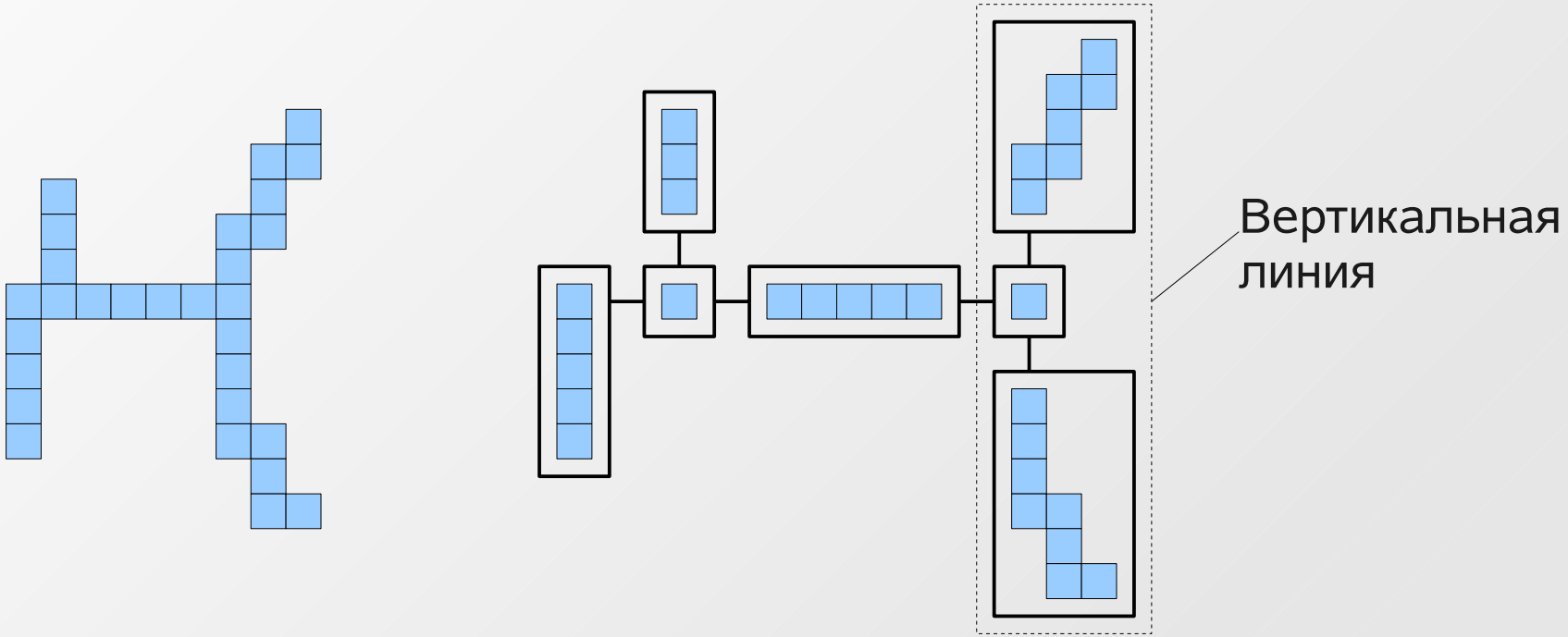
Форма линии



Положение точки



Трассировка линий



Формат запроса:

- Прямоугольная область поиска;
- Точка начала поиска;
- Направление поиска;
- Форма линии;
- Путь линии.

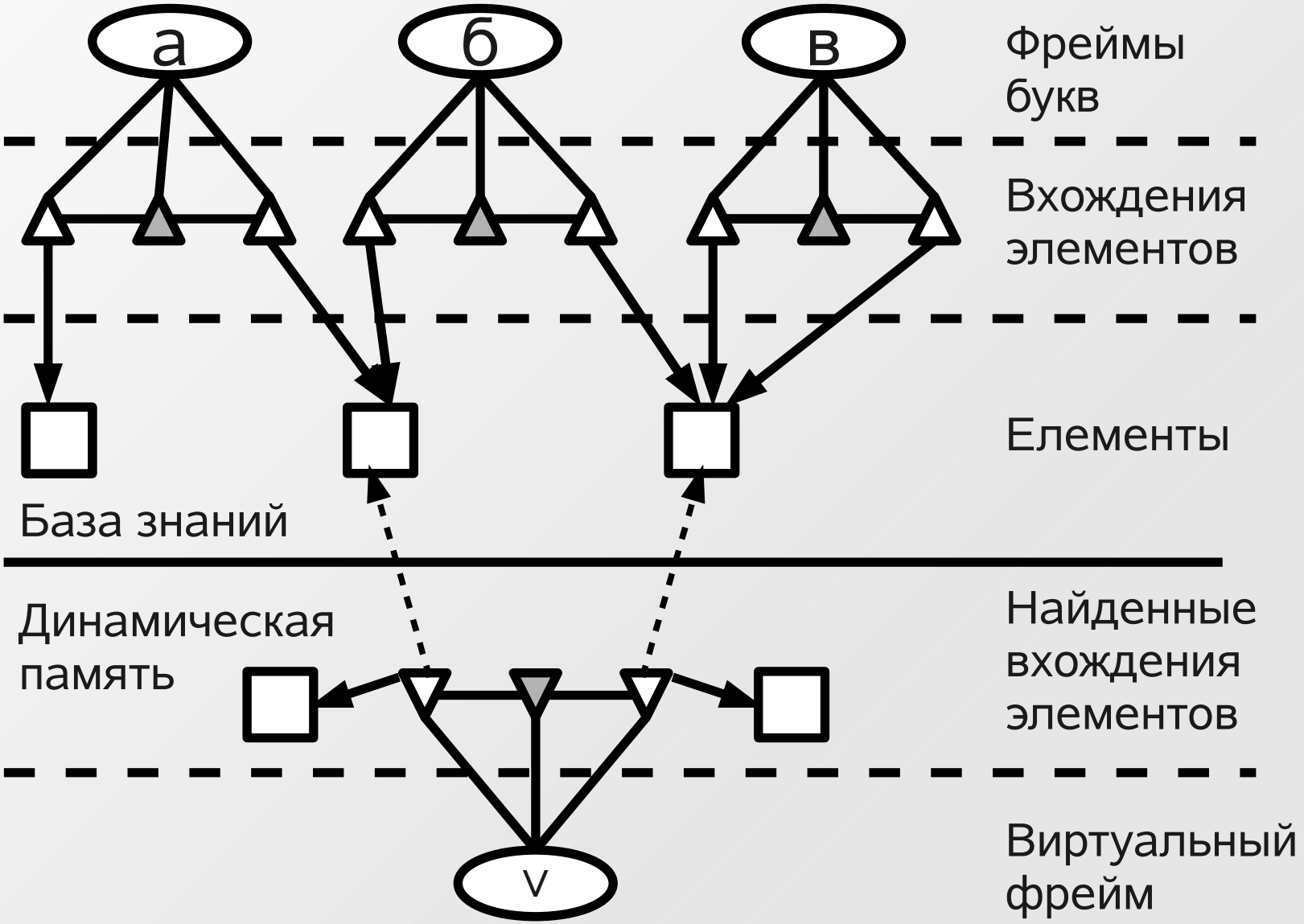
Формат ответа:

- Признак результата;
- Набор точек, составляющих найденную линию;
- Форма найденной линии;
- Путь найденной линии;
- Список точек, в которых данная линия пересекается другими.

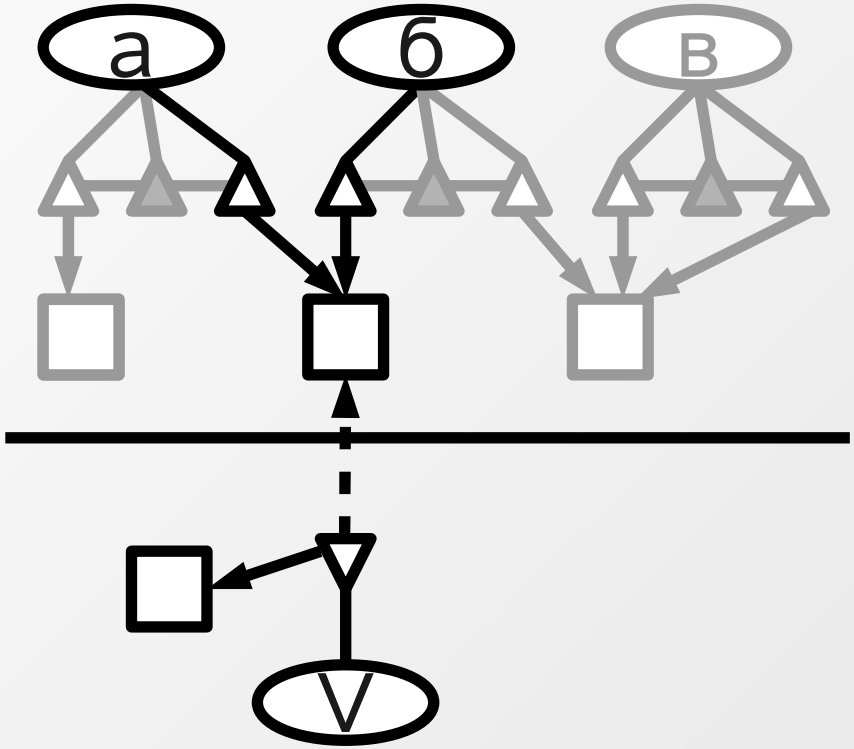
Статистика обучения системы

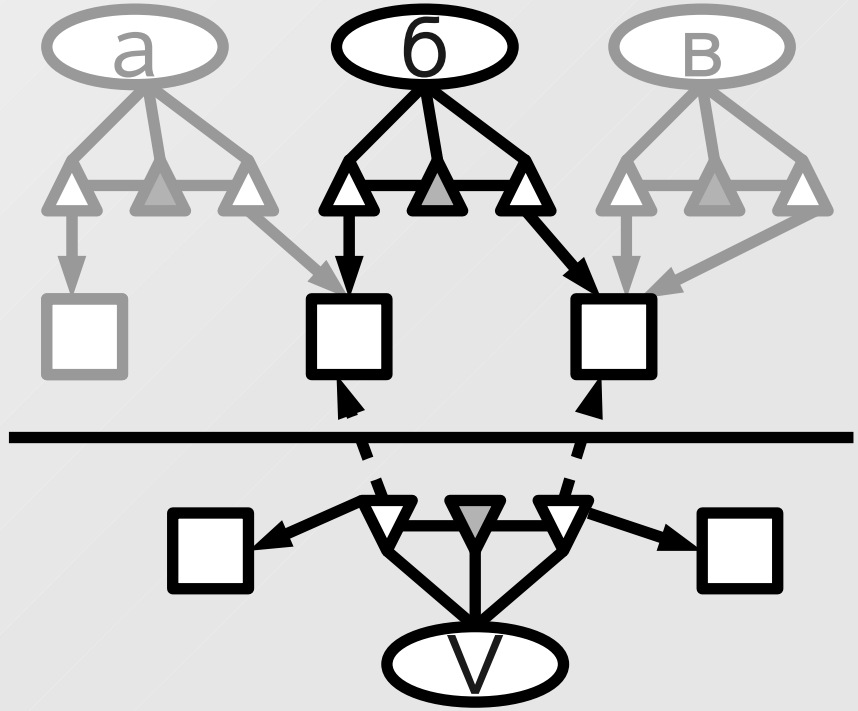
- Число букв — 38;
- Число различных начертаний — 300;
- Среднее число начертаний на букву — 7,89;
- Среднее число линий в начертании буквы — 2,3;
  - Из них различных — 2,07;
- Число типов линий — 89;
- Линия каждого типа встречается в среднем в 7,76 буквах;
  - Из них 6,98 различных.
- Размер файла OWL — 1,5 Мб;
- Размер базы знаний в памяти — 60 Мб.

Виртуальный фрейм

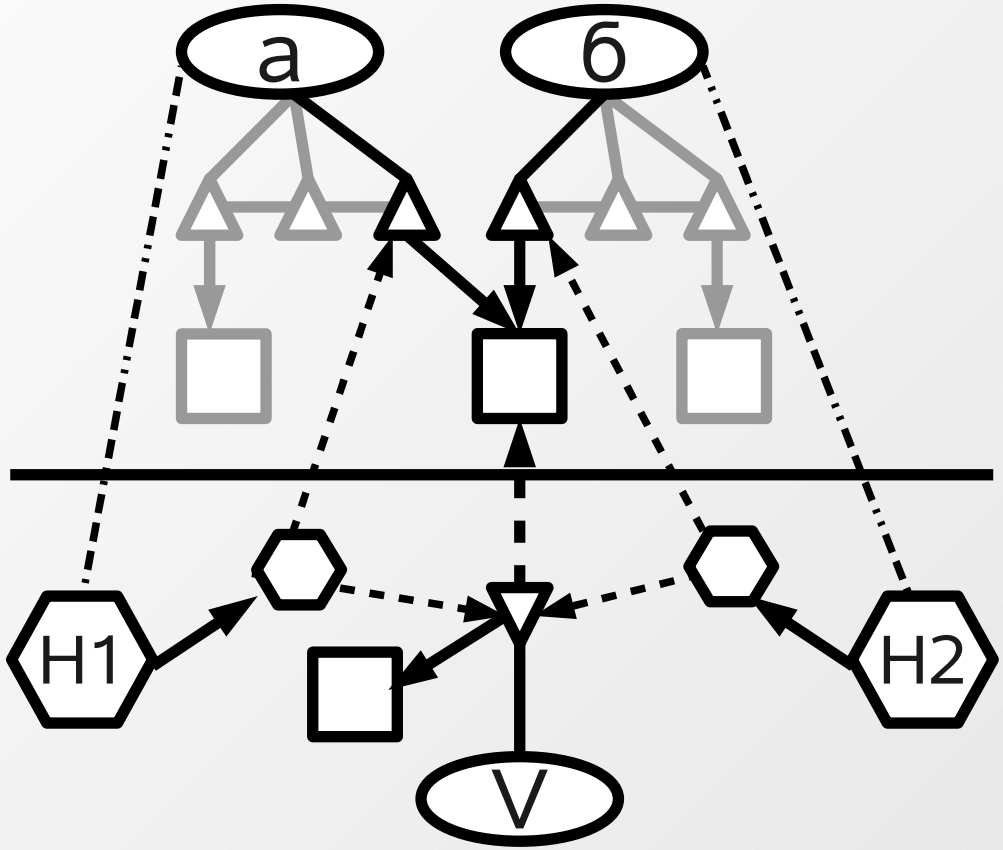


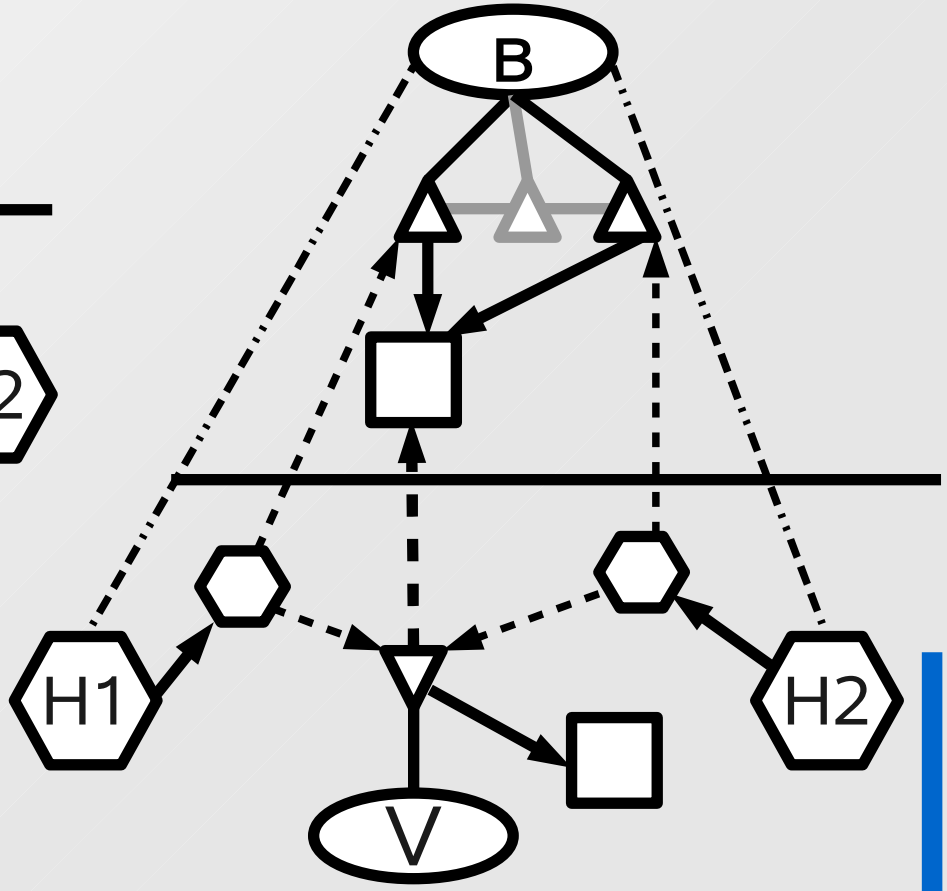
Выдвижение гипотез по найденным элементам





Построение гипотез





Характеристики гипотез и их проверка

- $Q = Q_{\ln} \cup Q_{PT}$ — множество *Вхождений свойств* в фрейме буквы;
- $Q_V = Q_{VLN} \cup Q_{VPT}$ — множество *Вхождений свойств* в ВФ;
- $H = \{(q \in Q, q_V \in Q_V)\} = H_{\ln} \cup H_{PT}$ — множество пар согласованных *Вхождений* из Q и Q_V в данной гипотезе;
- $w(Q) = |Q_{\ln}| * \alpha_{\ln} + |Q_{PT}| * \alpha_{PT}$ — вес набора *Вхождений свойств* в фрейме буквы или ВФ,

  где $\alpha_{\ln}, \alpha_{PT}$ — веса вхождений линий и букв соответственно;
- $w(H) = |H_{\ln}| * \alpha_{\ln} + |H_{PT}| * \alpha_{PT}$ — вес набора пар *Вхождений свойств* в гипотезе;
- $S_c = \frac{w(H)}{w(Q)}$ — *степень согласованности* гипотезы — показывает полноту соответствия ВФ фрейму буквы в соответствии с данной гипотезой;
- $S_a = \frac{w(H)}{w(Q_V)}$ — *степень пригодности* гипотезы — показывает точность соответствия ВФ фрейму буквы.

Программная реализация

- Языки реализации — Java, C++;
- Реализация базы знаний — язык OWL;
- Программные средства для манипулирования объектами базы знаний — библиотека Jena;
- Подсистема графического анализа — библиотека TINA.