Автоматизация распознавания древнерусских скорописных текстов

Цель:

- автоматизация перевода растровых изображений древнерусских скорописных текстов в электронное текстовое представление.

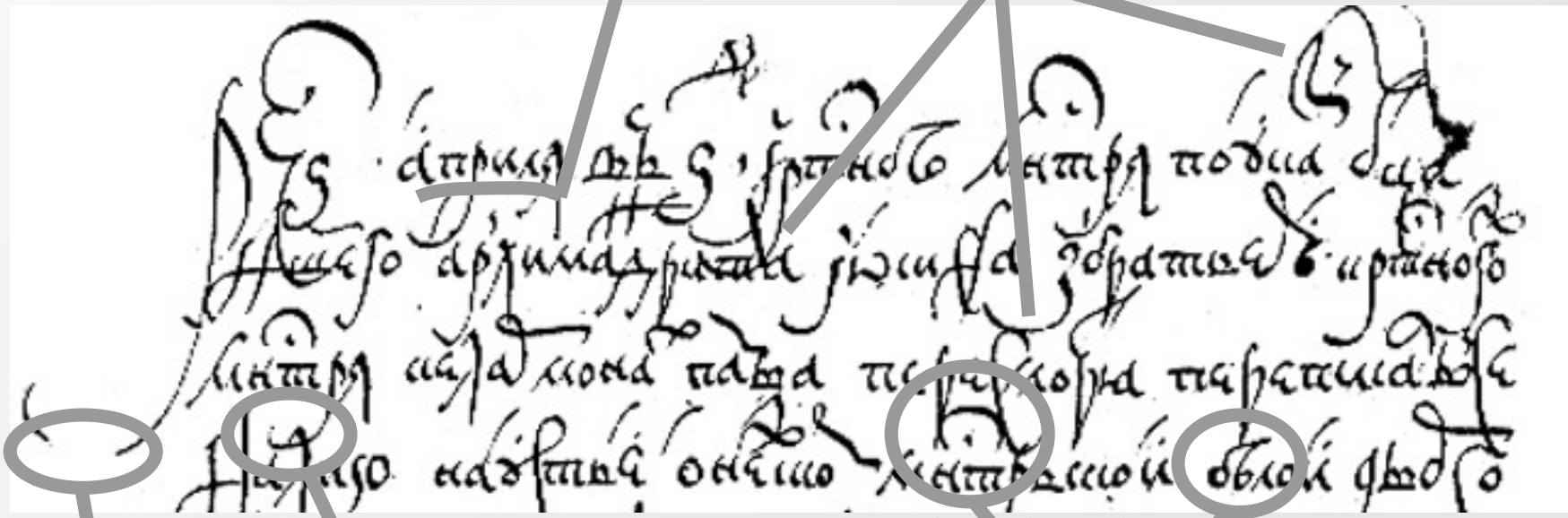
Задачи:

- изучение особенностей древнерусской скорописи;
- анализ существующих методов распознавания и разработка подходящего для решения задачи;
- проектирование и реализация системы распознавания.

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

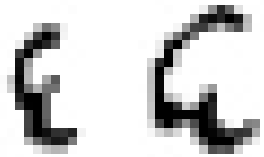
Декоративные росчерки

Дефект

Пересечение элементов букв

Соединение букв

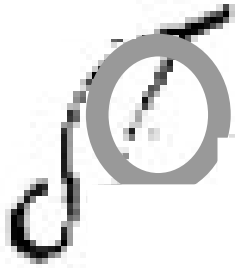
Особенности скорописного способа формирования текста



Варьируемость начертания

Отсутствует пересечение

Лишнее пересечение

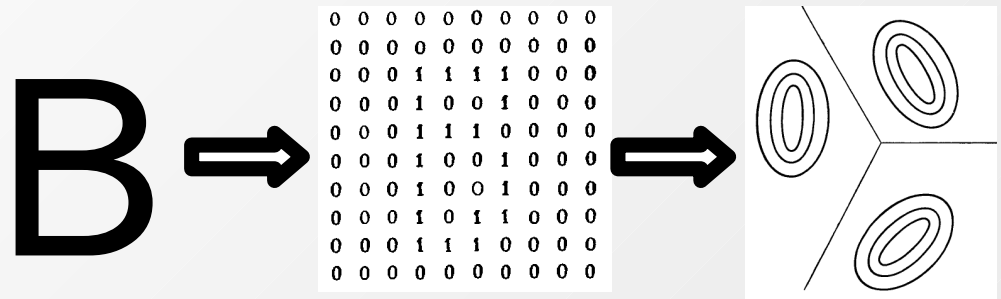


Декоративный элемент

Дефекты

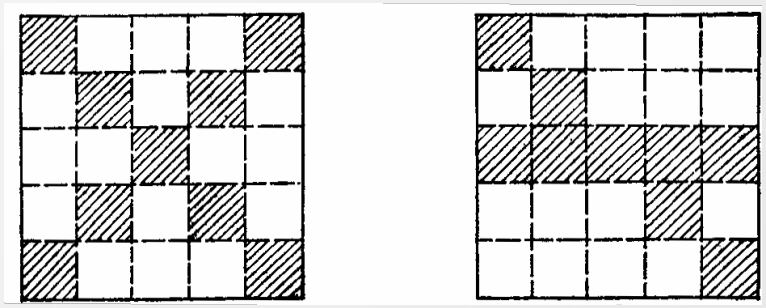
Методы распознавания

Евклидово пространство



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

Признаковые методы



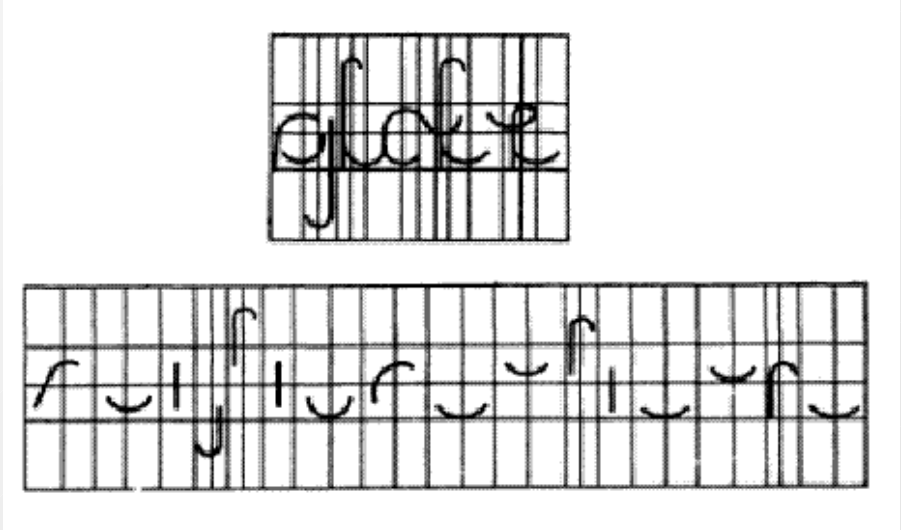
$$x_i = \{x_{i1}, x_{i2}, \ldots, x_{ik}\}$$

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.

Методы распознавания

Структурные методы

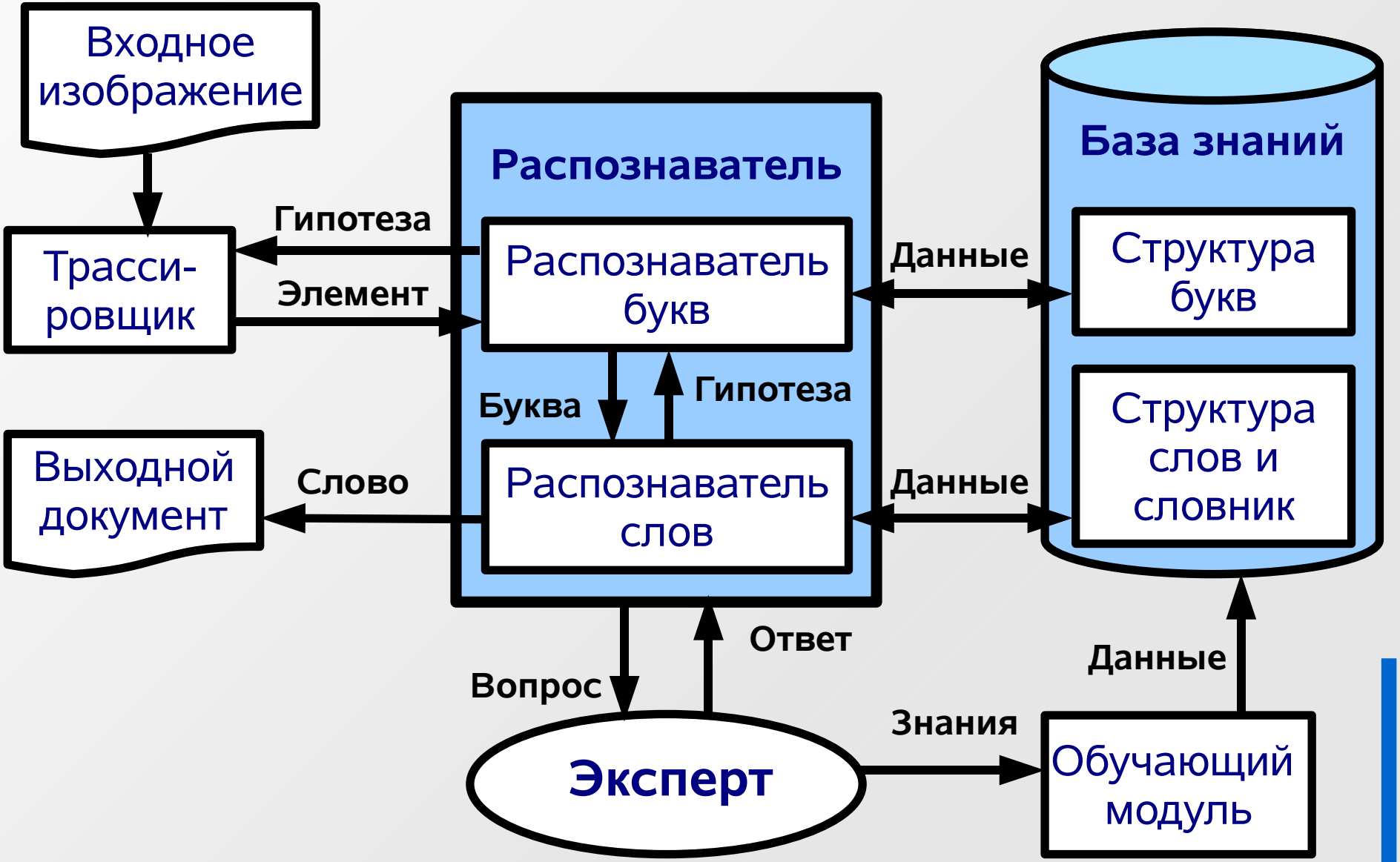


- основываются на выделении и анализе составных частей распознаваемых объектов;
- позволяют целенаправленно перемещаться по изображению в процессе распознавания;
- позволяют учитывать декоративные элементы.

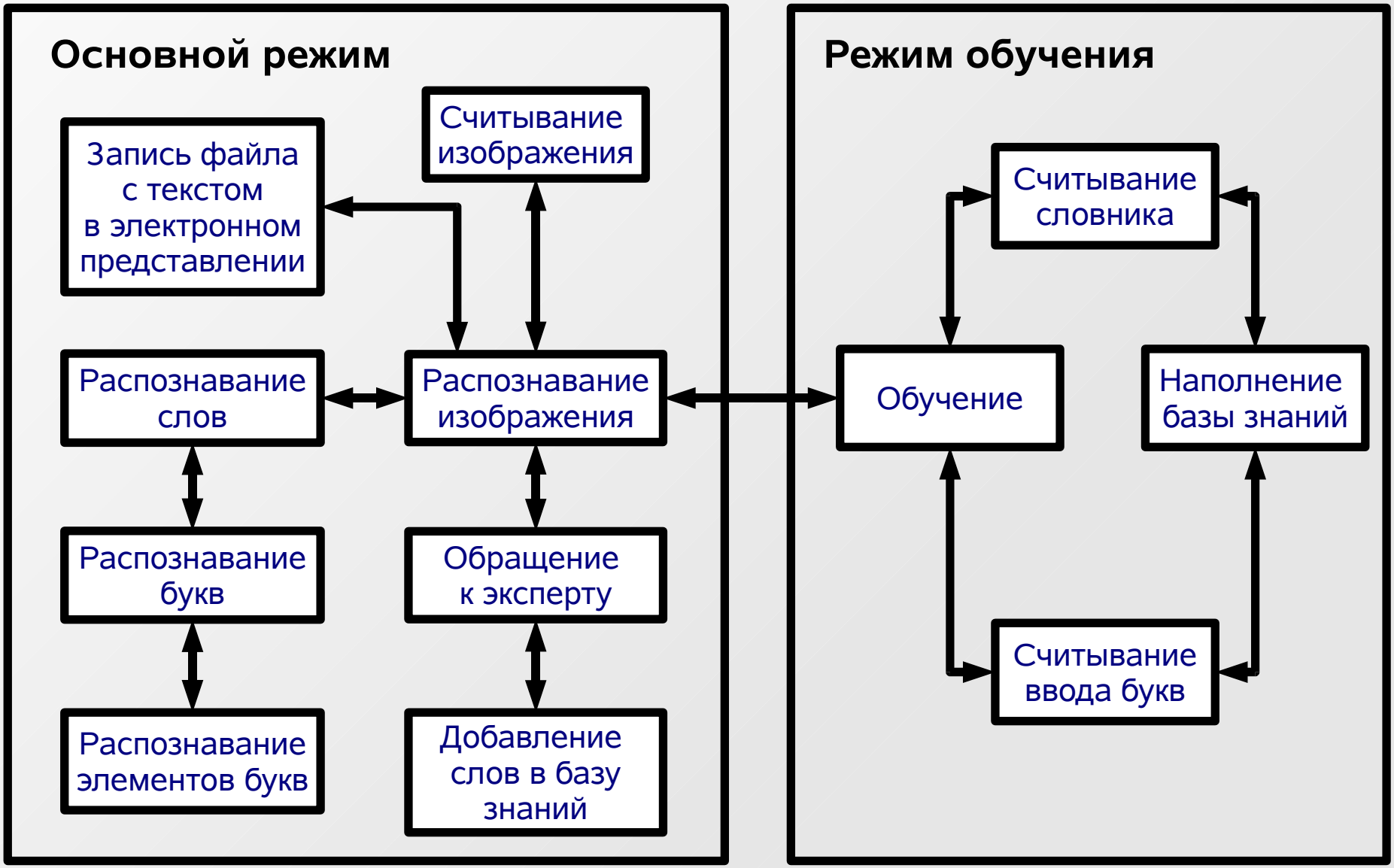
Ключевые особенности разработанного метода распознавания

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

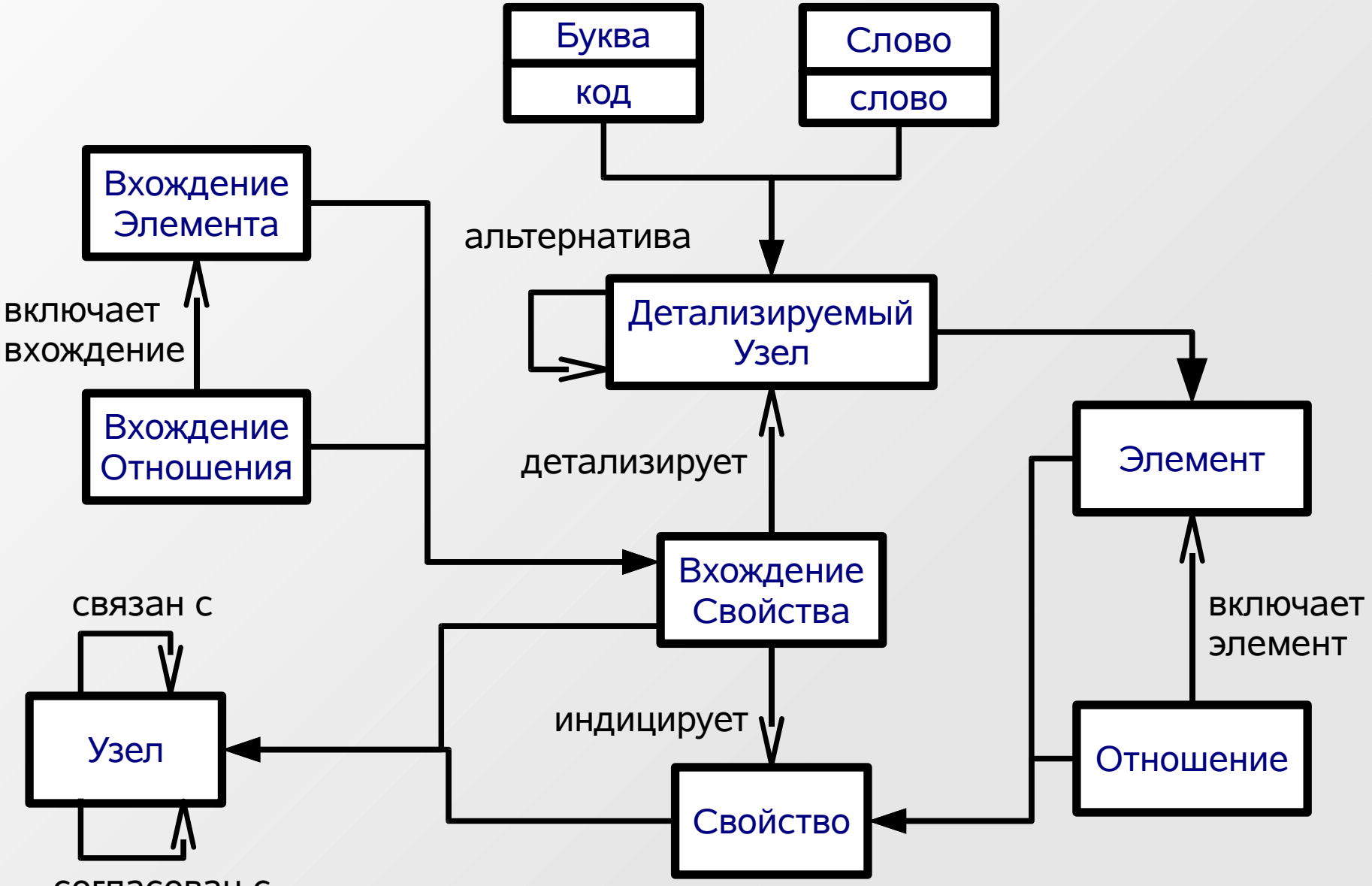
Система распознавания. Структурная схема



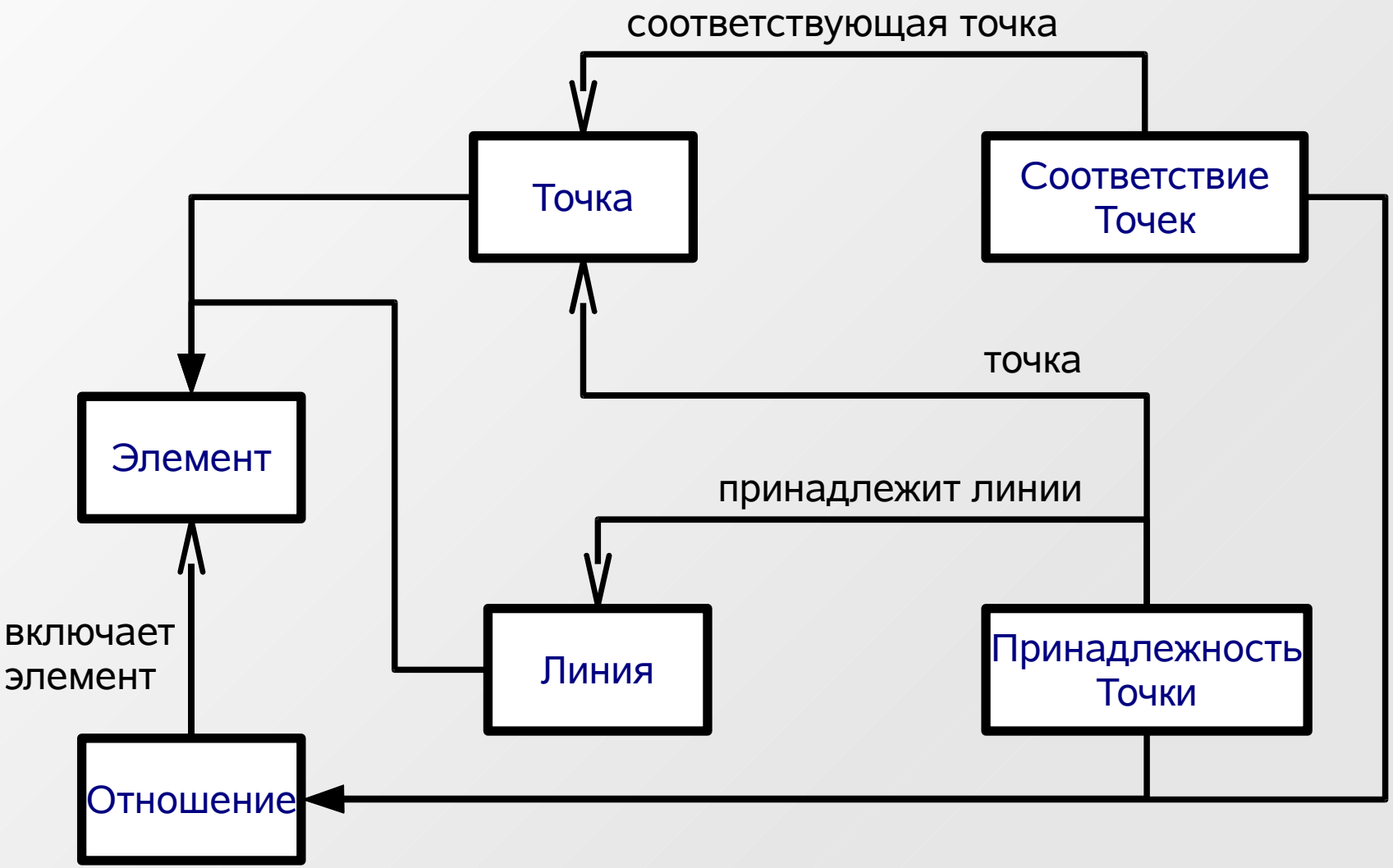
Система распознавания. Функциональная схема



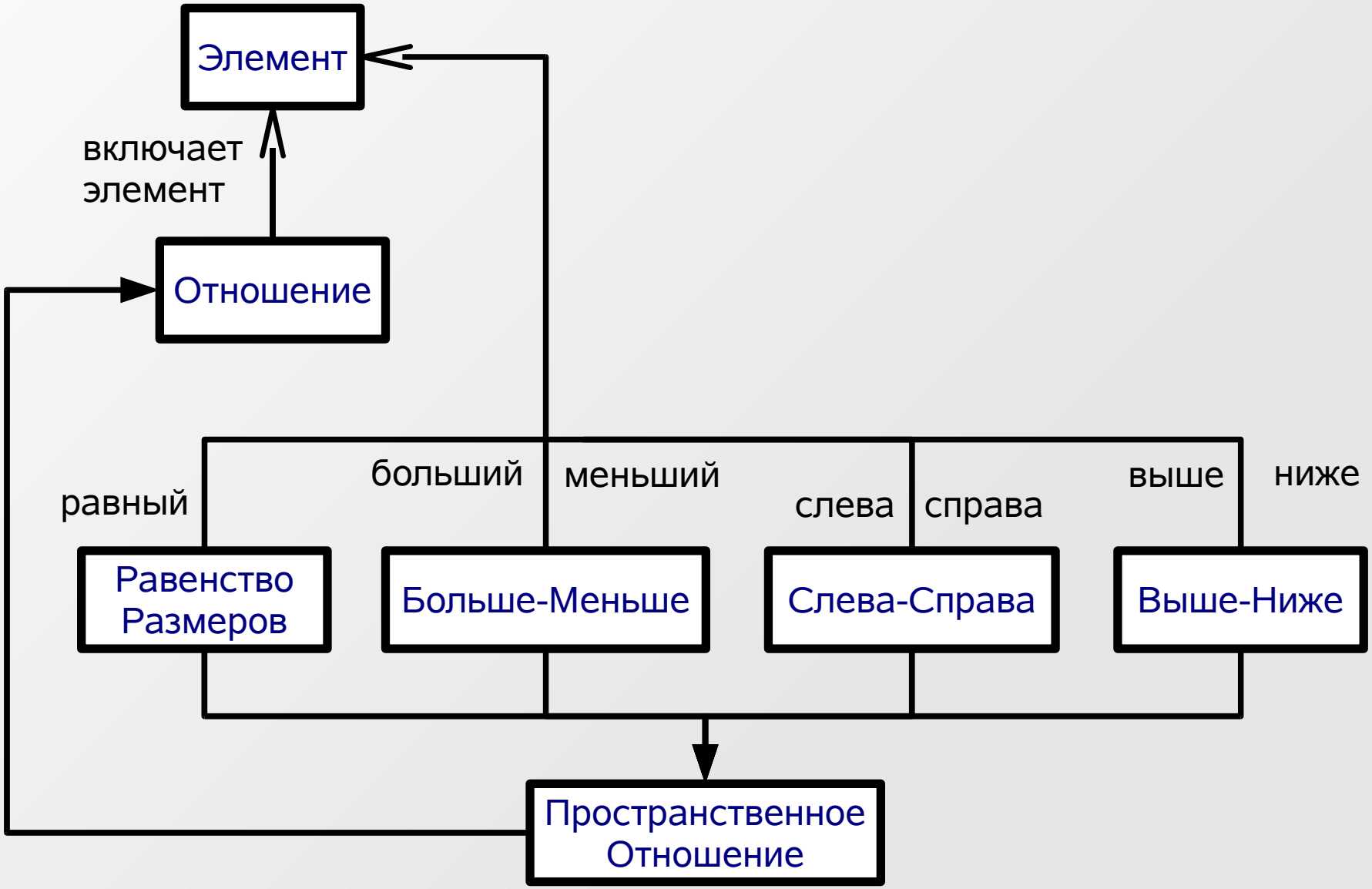
Структура базы знаний



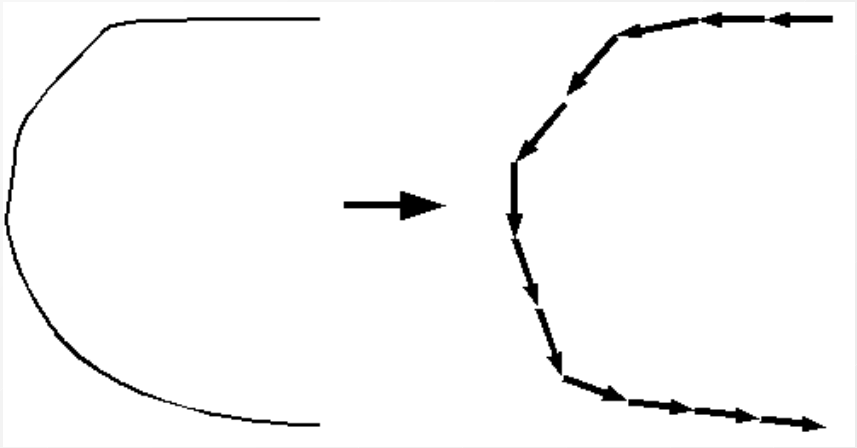
Структура базы знаний

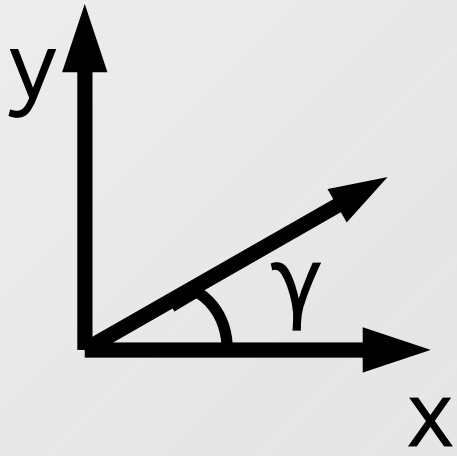


Структура базы знаний

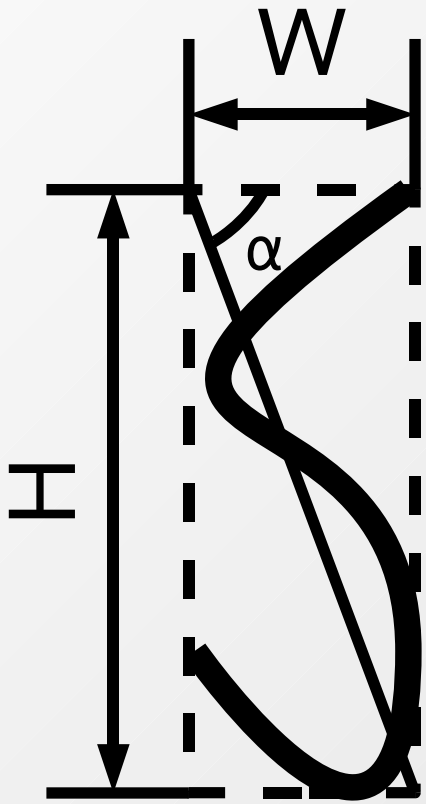


Путь линии

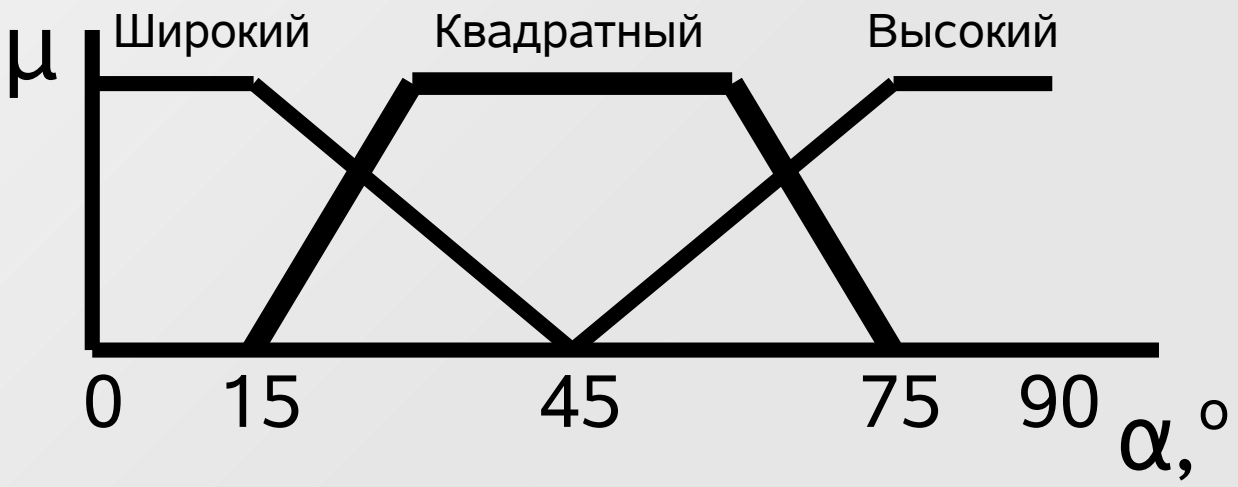




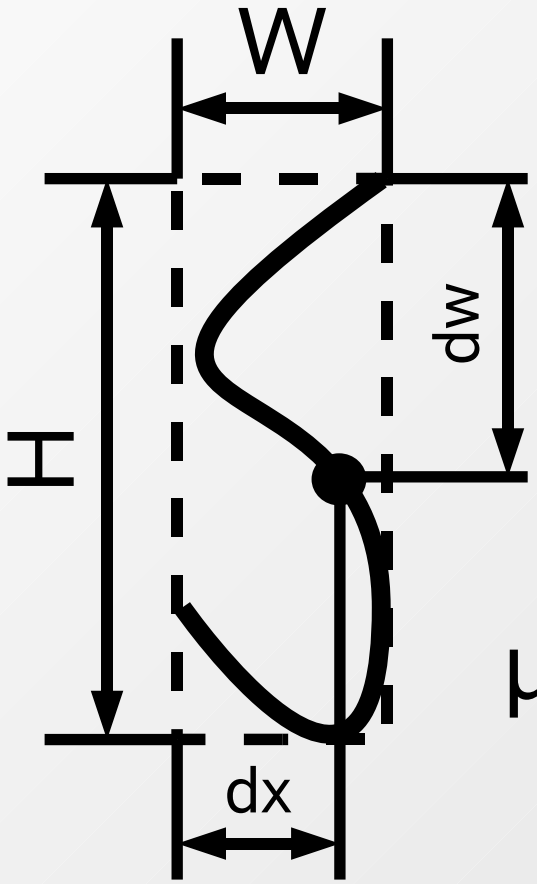
Форма линии



$$\alpha = arctg\left(\frac{H}{W}\right)$$

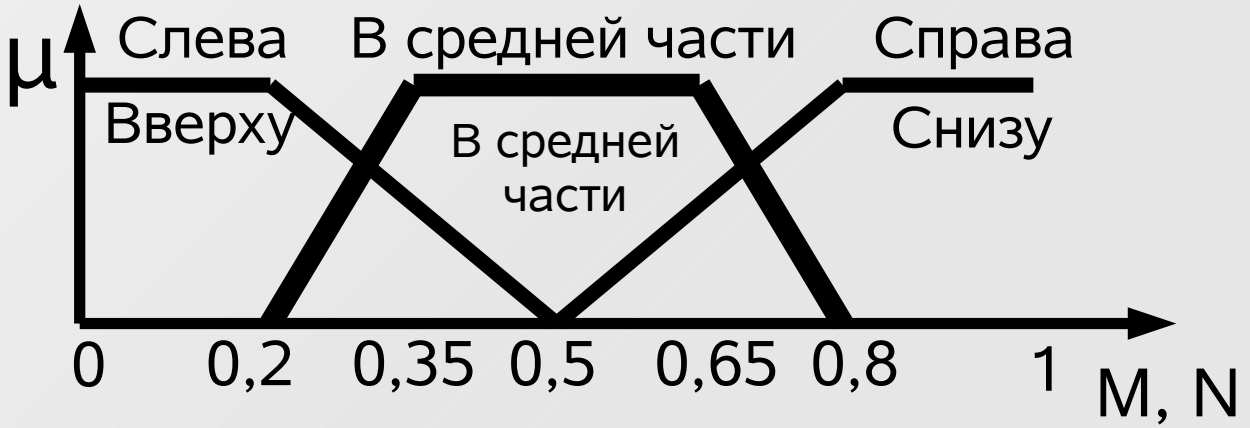


Положение точки

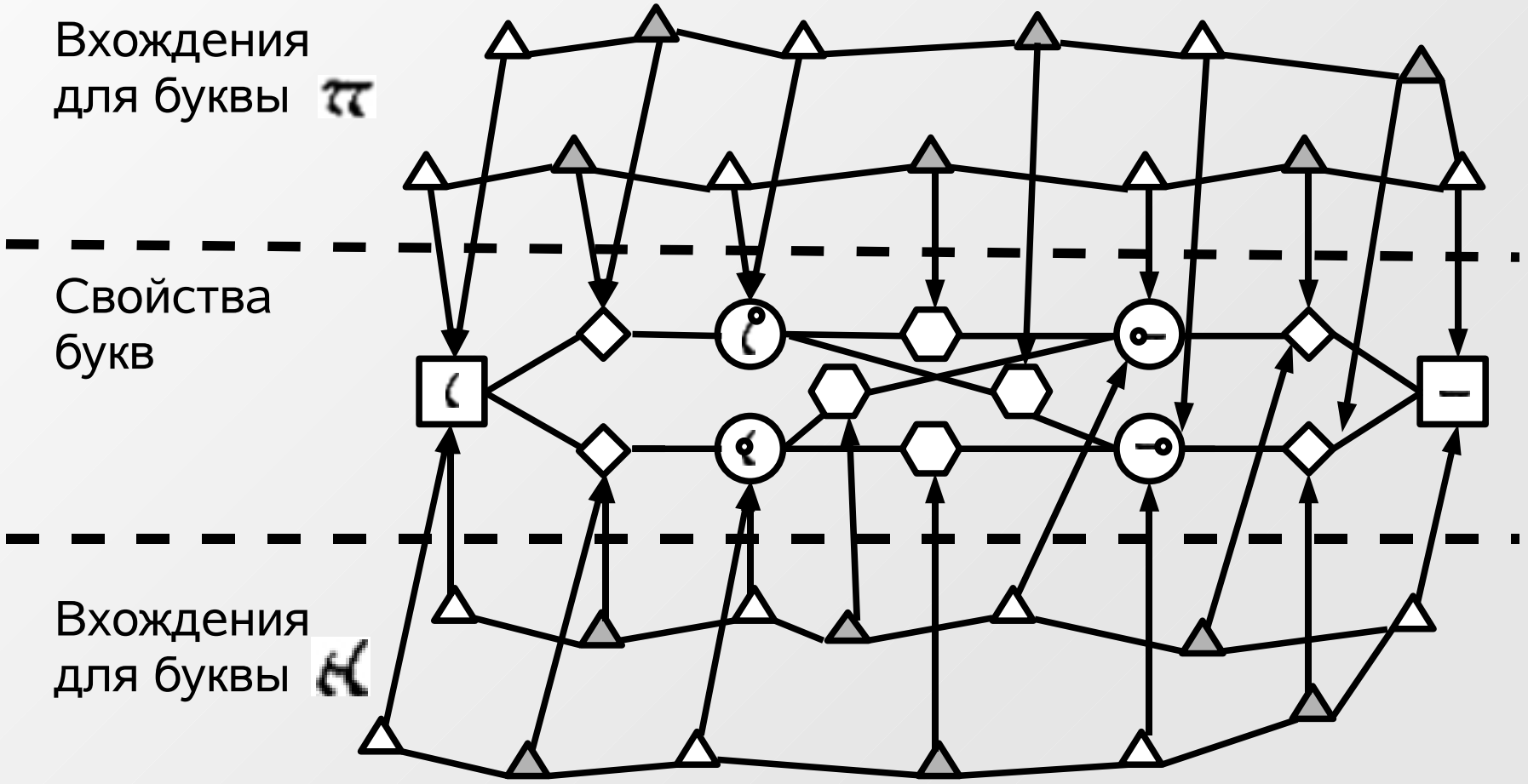


$$M = \frac{dw}{W}$$

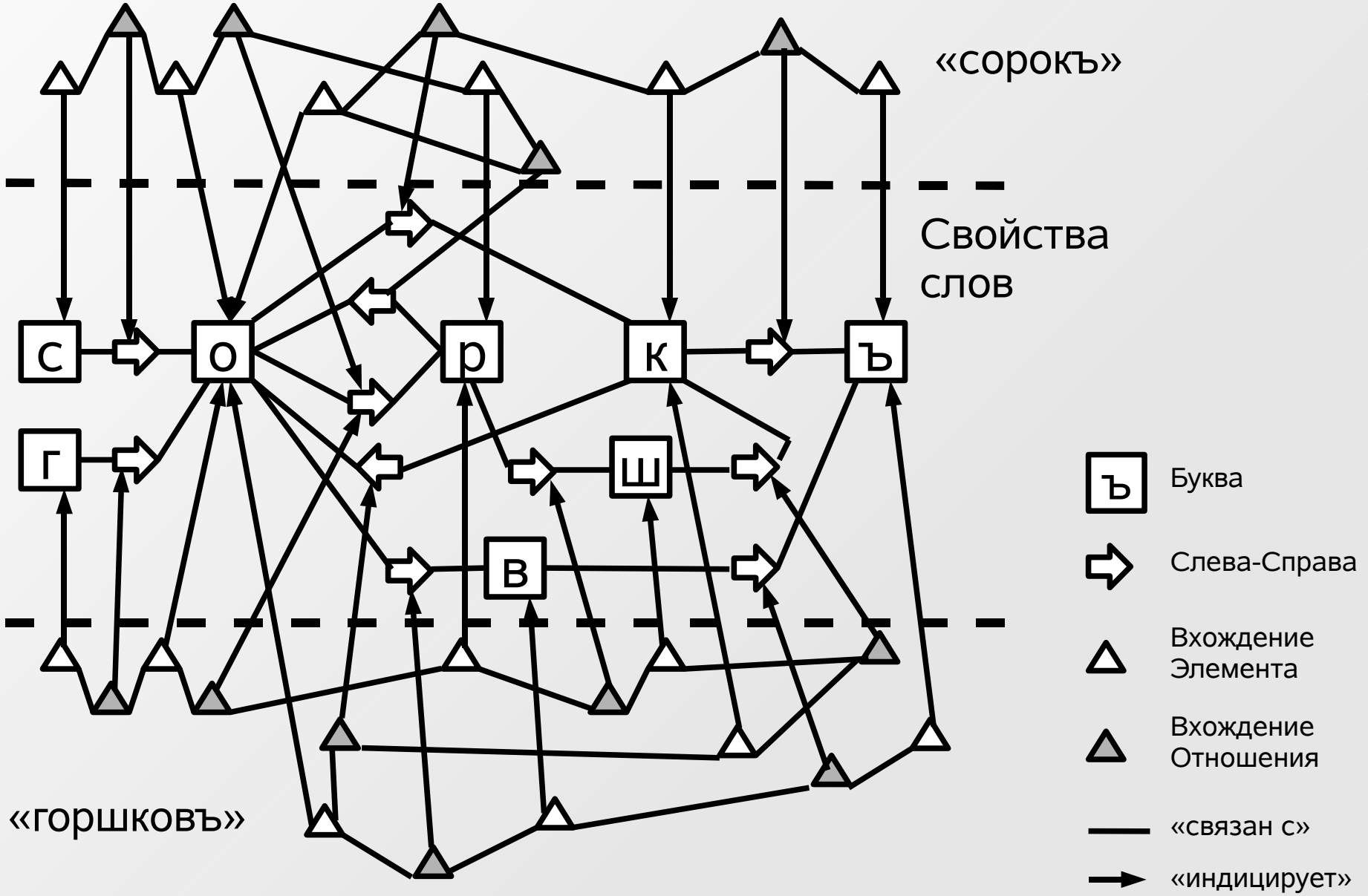
$$N = \frac{dh}{H}$$



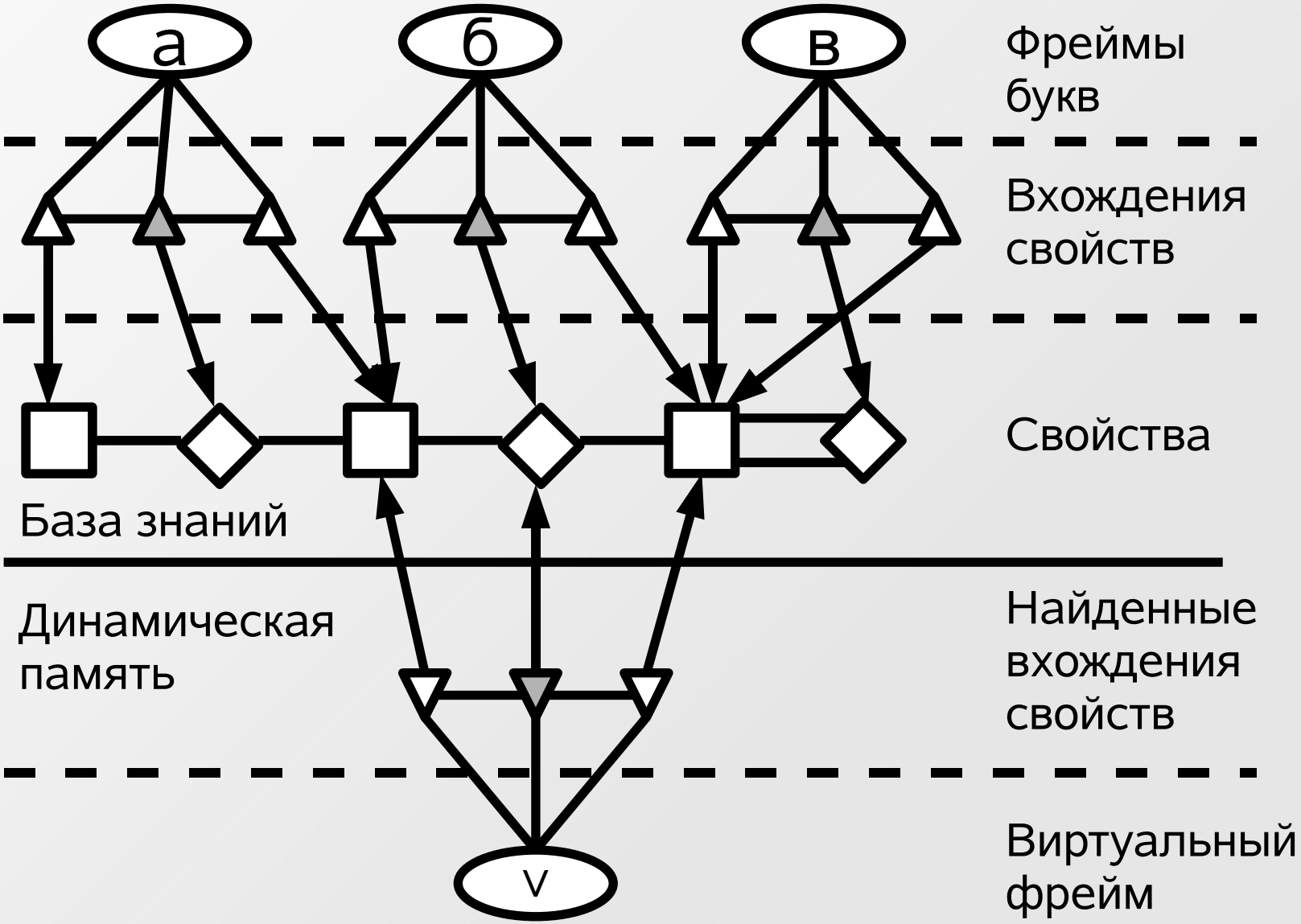
Пример фреймов букв



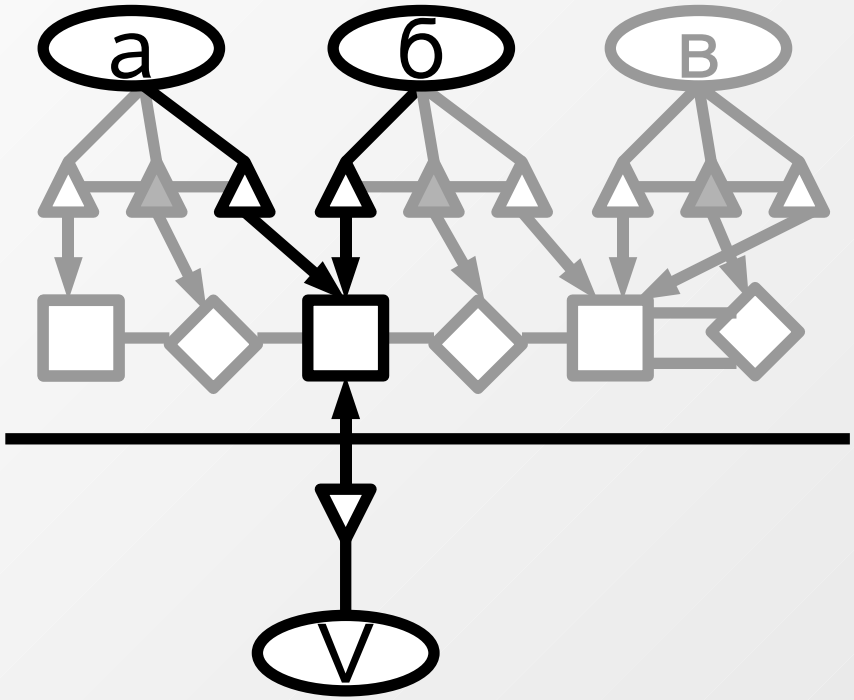
Пример фреймов слов

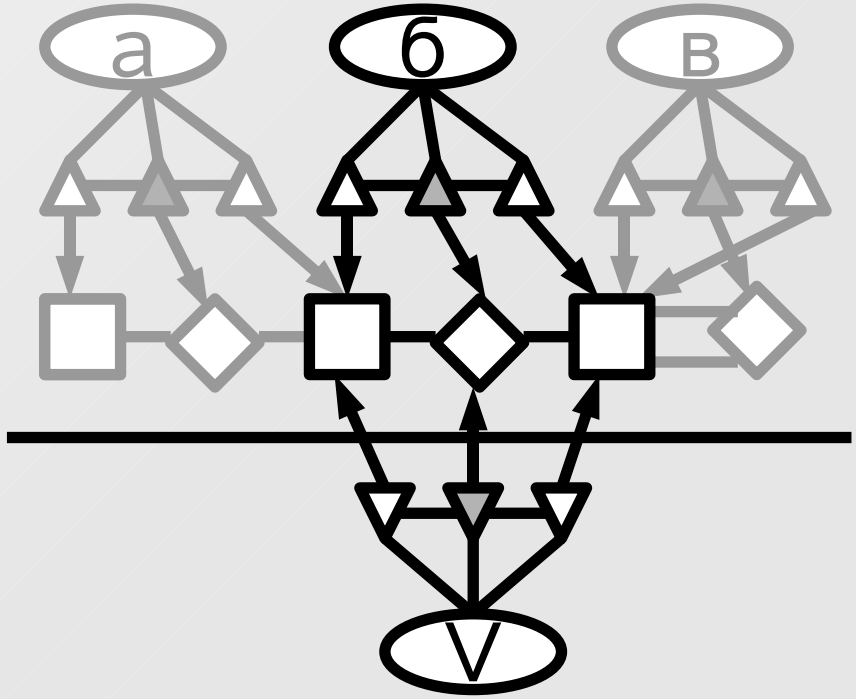


Виртуальный фрейм

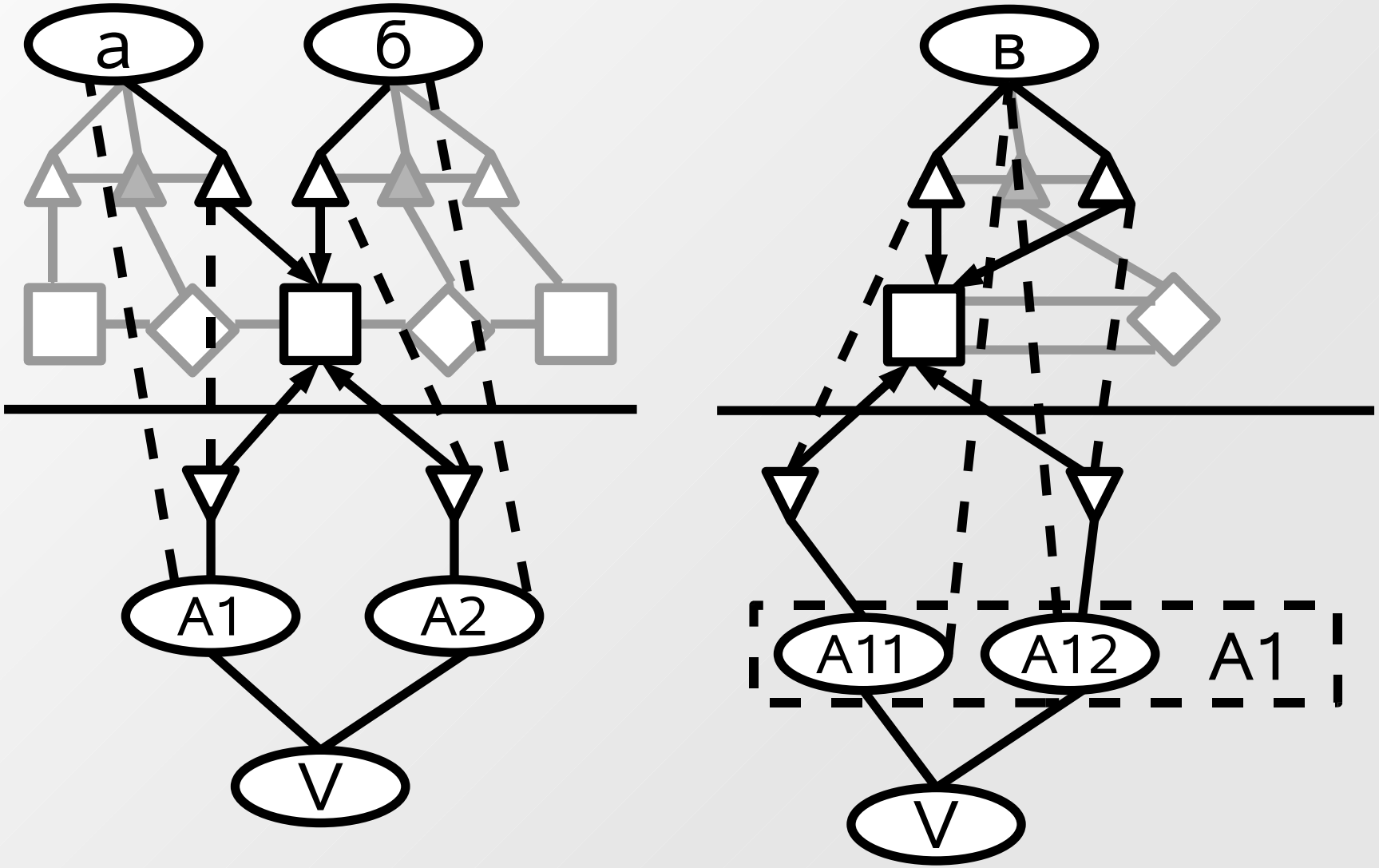


Выдвижение гипотез по найденным элементам





АВФ: построение фреймов гипотез



Характеристики АВФ и проверка гипотез

- Q — множество *Вхождений свойств* в фрейме буквы;
- V — множество *Вхождений свойств* в АВФ;
- N — число пар согласованных *Вхождений* из Q и V;

- $S_c = \frac{N}{|Q|}$ - *степень согласованности* АВФ — показывает полноту АВФ соответствует фрейму-гипотезе;
- $S_a = \frac{N}{|V|}$ - *степень пригодности* АВФ — показывает точность соответствия АВФ фрейму буквы;

- $S_c > П_c$ - *условие согласованности* — говорит о подтверждении гипотезы, описываемой данным АВФ;
- $S_a > П_a$ - *условие пригодности* — является необходимым условием для продолжения проверки гипотезы, описываемой данным АВФ.

Программная реализация

- Язык реализации — Java;
- Реализация базы знаний — язык OWL;
- Программные средства для манипулирования объектами базы знаний — библиотека Jena;
- Средства долговременного хранения базы знаний — СУБД Apache Derby, файлы OWL в нотации XML;
- Подсистема графического анализа — библиотека TINA.