

# Объектная модель многофункциональных словарей

Докладчик: Носков А. А.

Группа: 525

Научный руководитель: Большакова Е. И.

# Рассматриваемая работа

«Объектная модель многофункциональных словарей, основанная на синтезе лингвистических единиц»

Ивличева О. О., Епифанов М.Е., Лахути Д.Г.

Попытка выработать универсальный метод организации данных для электронных словарей

# Лингвистические единицы

- Морфема, лексема, словоформа...
- Простая/составная: морфема/словоформа
- Многоуровневая иерархия
  - Синтаксические конструкции образованы из словоформ, словоформы из основы и флексий
- Составная л.е. - результат некоторой операции над единицами нижнего уровня
  - Словоформа может быть получена как конкатенация ее составляющих

# Свойства лингвистических единиц

- С единицами ассоциированы некоторые свойства
- Внутренние/наследуемые (для составных частей)
  - Словосочетание «большой корабль» наследует свойства рода, одушевленности и т.п. от «корабль»
  - Можно считать, что словоформа наследует свой падеж от окончания
- Со значением некоторых свойств связано «поведение», в частности, правила построения новых единиц
  - Род, число определяют согласование

# Текст-объект

- В словаре конкретные лингвистические единицы представляются в виде текст-объектов
- Основы: «вершин», «дорог»
- Окончания: «а», «и», «ы»...
- **Текст-объект** — цепочка символов + конечное множество свойств

# Свойства текст-объекта

- **Свойство** — тройка  $p = \langle d, n, v \rangle$ , где
  - $d$  — тип свойства
  - $n$  — имя свойства
  - $v$  — значение свойства
- У одного текст-объекта не может быть свойств с одинаковым именем
- Примеры свойств
  - Падеж, число, одушевленность и прочие грамматические признаки
  - Семантическое значение суффикса

# Аддитивные и внутренние свойства

- Свойства делятся на *аддитивные* и *внутренние*
  - **Аддитивные** — свойства, которые наследуются более сложными конструкциями
    - Род, число, одушевленность
  - **Внутренние** — свойства, которые не наследуются
    - Тип единицы
    - Часть речи
- Множества имен аддитивных и внутренних свойств не пересекаются
- Текст-объект - тройка  $\langle t, AData, IData \rangle$  (строка, аддитивные свойства, внутренние свойства)

# Соединение текст-объектов

- Используется для образования составных текст-объектов из более простых
- Текст-объекты *соединимы*, если все их аддитивные свойства могут быть успешно соединены
- <«**вершин**» , { **одуш:неод** } , ∅> соединима с <«**ы**» , { **одуш:неод** , **число:ед** , **пад:вин** } , ∅> но не соединима с <«**»** , { **одуш:од** , **число:ед** , **пад:вин** } , ∅>



# Соединение свойств

- Для каждого типа свойства определяется специальный оператор соединения свойств простых текст-объектов (пары объектов) в свойства составного текст-объекта
- Тип «согласуемое свойство» переносит в новый текст-объект свойства, только если  $e_1$  и  $e_2$  не содержат одноименных свойств с различными значениями

# R-объекты

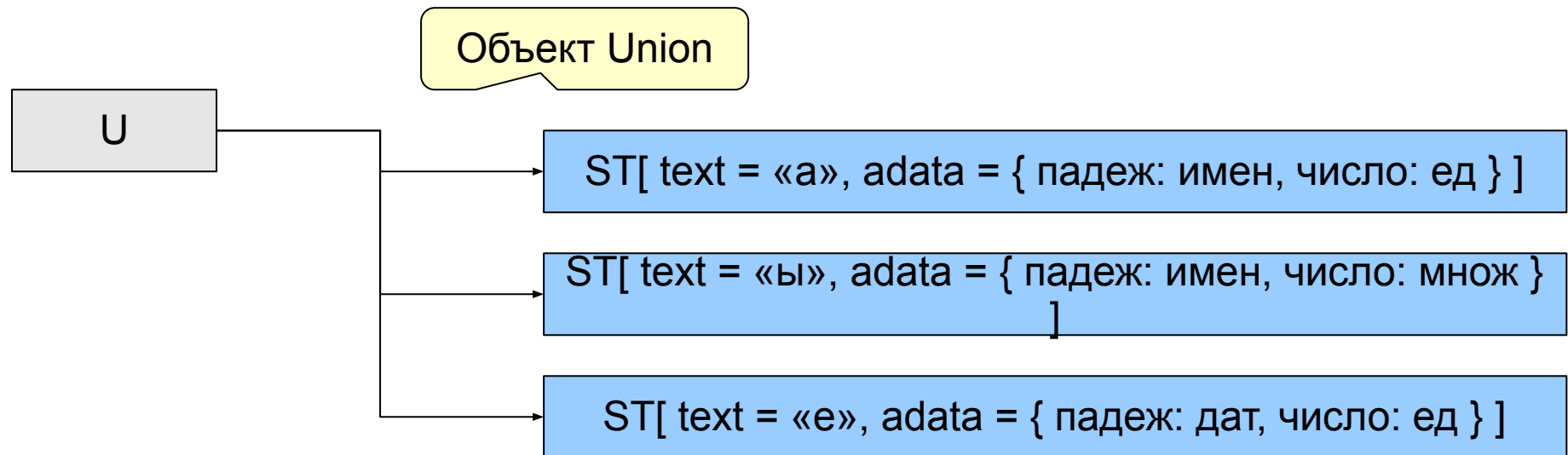
- R-объекты — собственно элементы структуры словаря.
- Могут быть четырех типов: SimpleText, Property, Union, Join
- Каждый R-объект  $e$  описывает какое-то множество текст-объектов
- Есть R-объекты, представляющие
  - Конкретные основы и флексии
  - Множества возможных основ, флексий
  - Множества допустимых словоформ
  - Множества допустимых словосочетаний

# R-объекты: SimpleText

- `ST [text, adata, idata]` — R-объект, соответствующий одному текст-объекту
- Ими представляются основы и флексии
  - `ST [«вершин», {одуш:неодуш}, ∅]`
  - `ST [«а», {род:жен, числ:ед, пад:им}, ∅]`
  - `ST [«ы», {одуш:неодуш, числ:мн, пад:вин}, ∅]`
  - `ST [«», {одуш:одуш, числ:мн, пад:вин}, ∅]`
- Property эквивалентен SimpleText без поля text

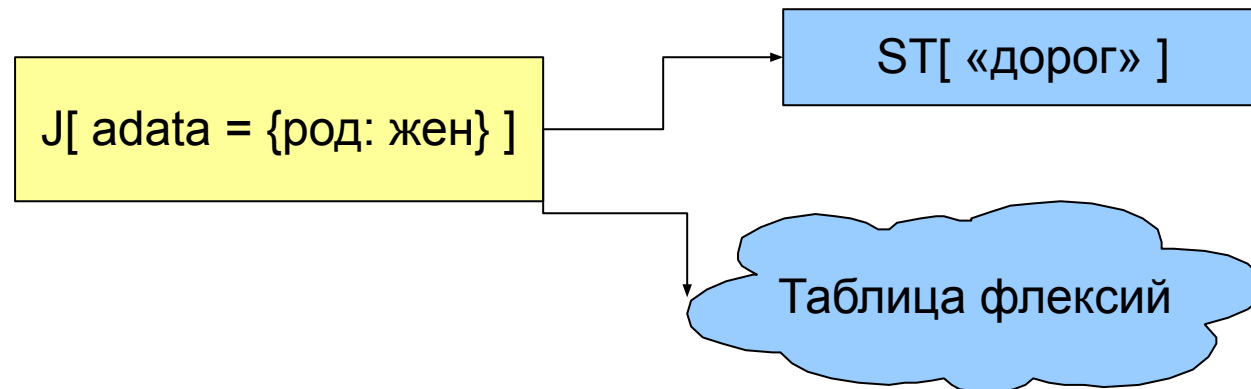
# R-объекты: Union

- Union — составной R-объект, который используется для объединения множеств, описываемых дочерними R-объектами
- Например, объект, объединяющий окончания в множество (таблицу флексий)



# R-объекты: Join

- Для представления множеств составных единиц используется объект Join
- Join представляет множество соединений всех пар дочерних объектов
- Им представляются множества словоформ, словосочетаний

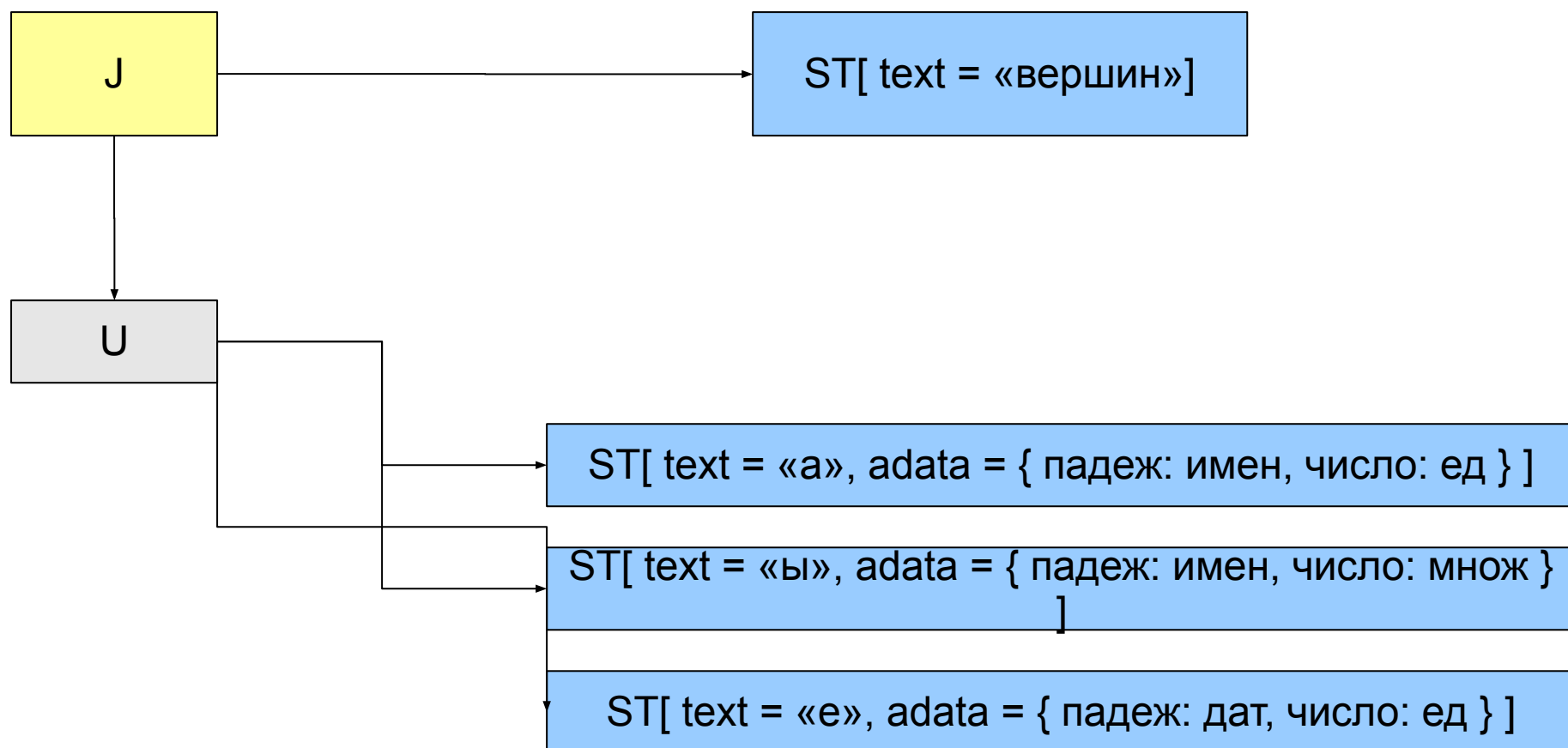


# Построение словаря из R-объектов

- R-объекты организованы в иерархию
- Листьями в иерархии являются минимальные единицы: морфы, представленные SimpleText
- Составными элементами являются Union и Join, ссылающиеся на другие R-объекты
- При применении Join к основе и множеству Union окончаний, основа «склеивается» с каждым окончанием

# Пример фрагмента словаря

Структура, описывающая слова «вершина», «вершины» и «вершине»



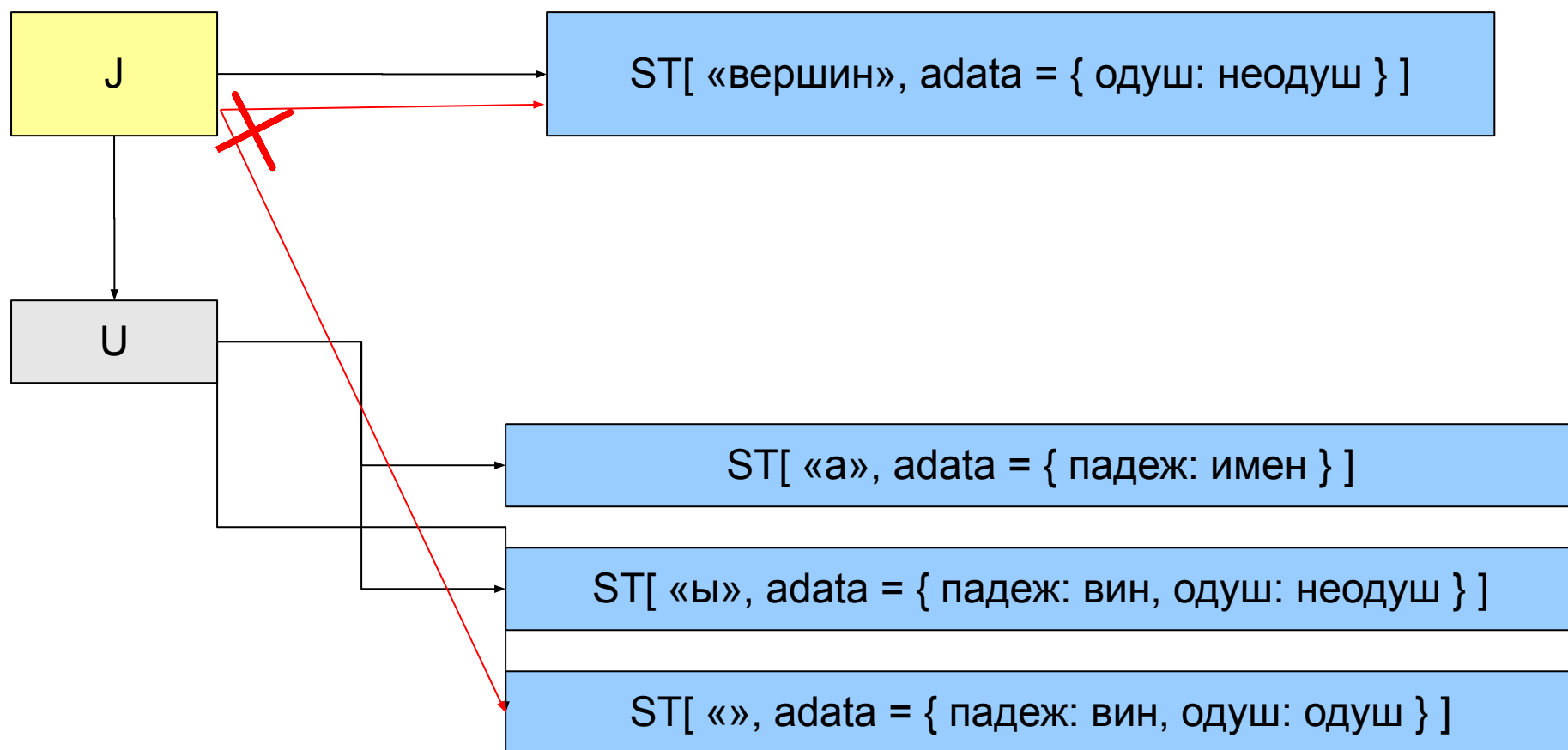
# Недопустимые единицы

- Соединение может отвергать некоторые единицы, получаемые в результате синтеза как неправильные.
- Такие единицы состоят из несоединяемых объектов и не входят в результирующее множество.
- Простейший пример: конфликт значений свойств.

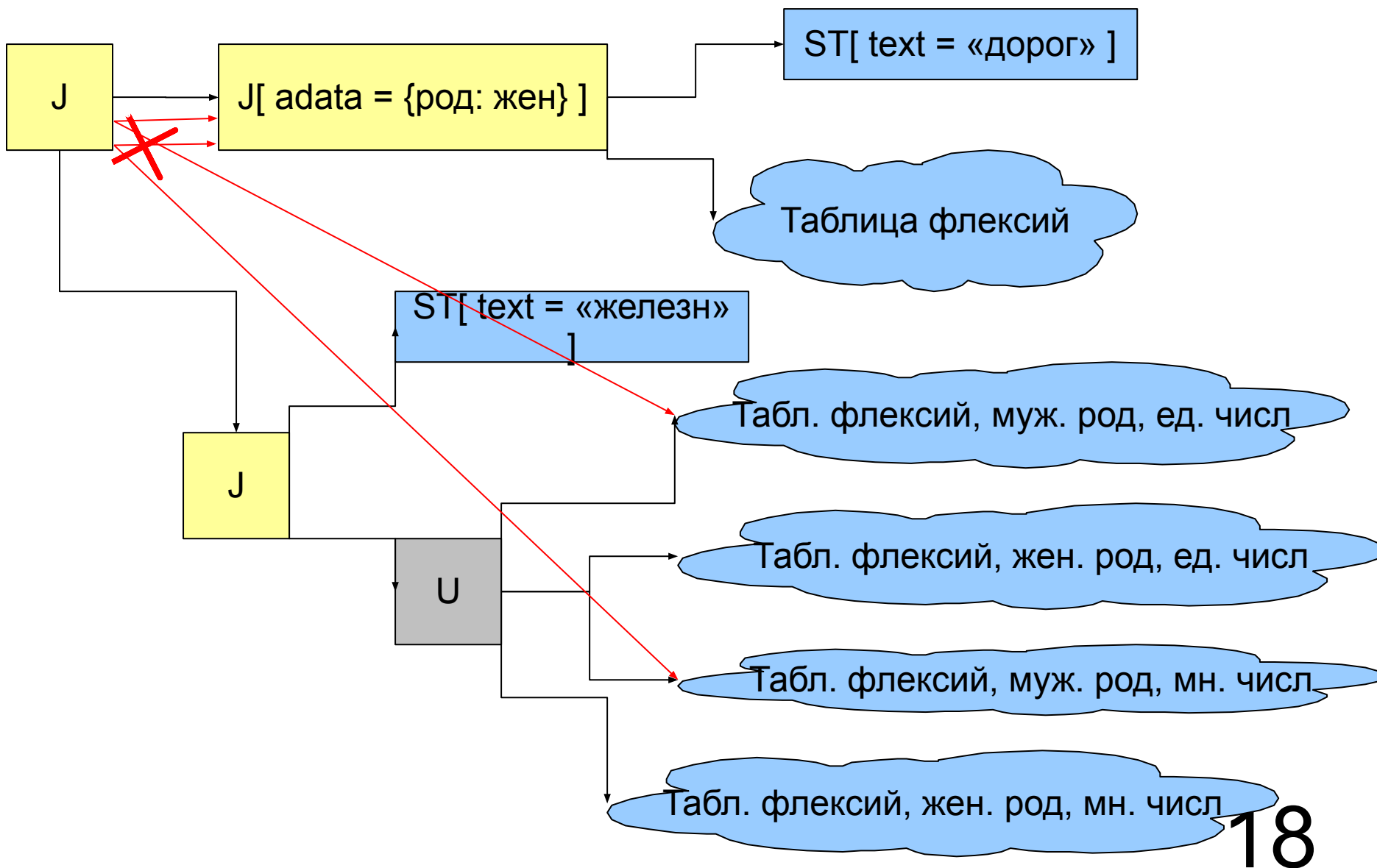


# Пример недопустимых единиц

Конфликт свойства одушевленности, «вершин» - недопустимая форма!



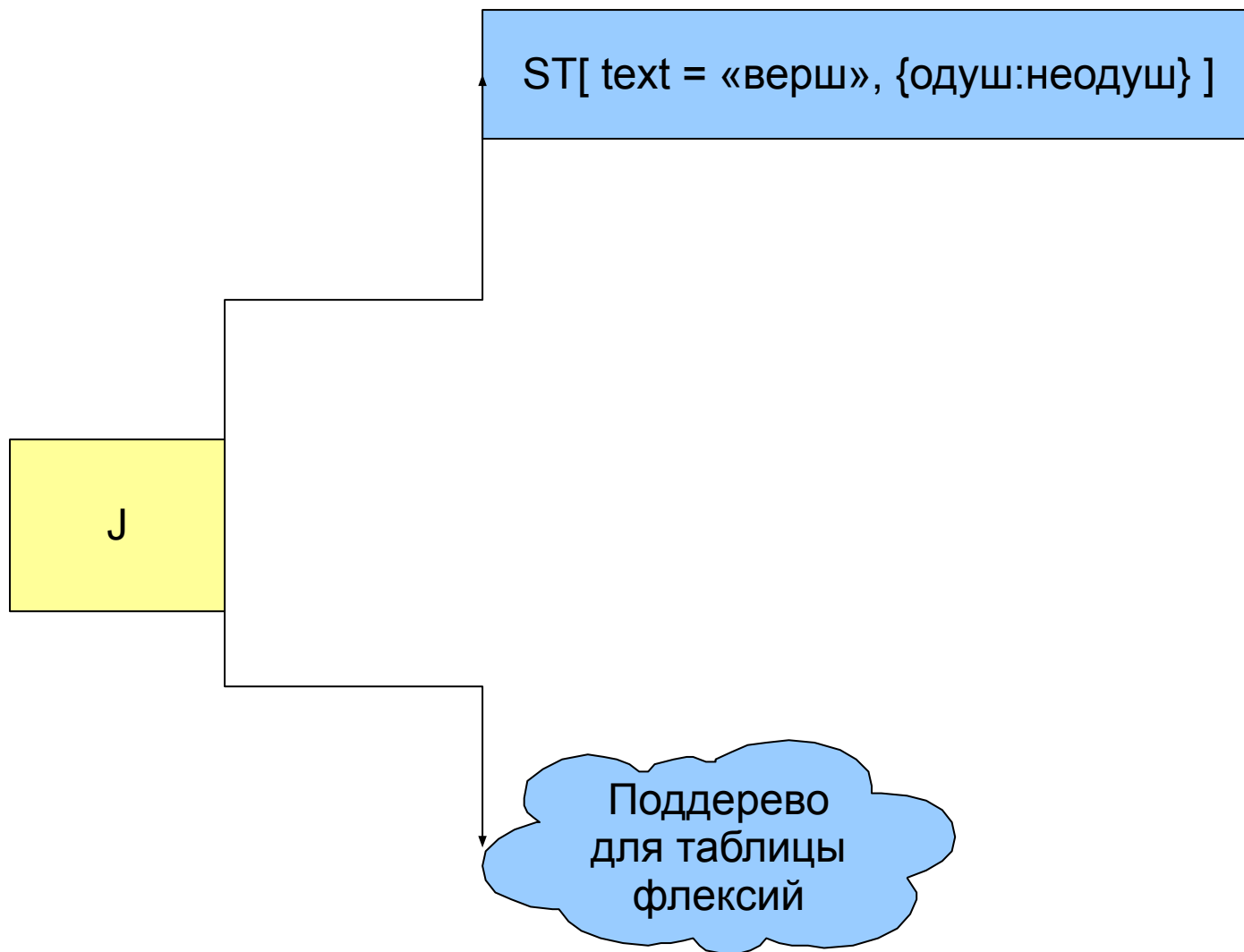
# Представление словосочетаний



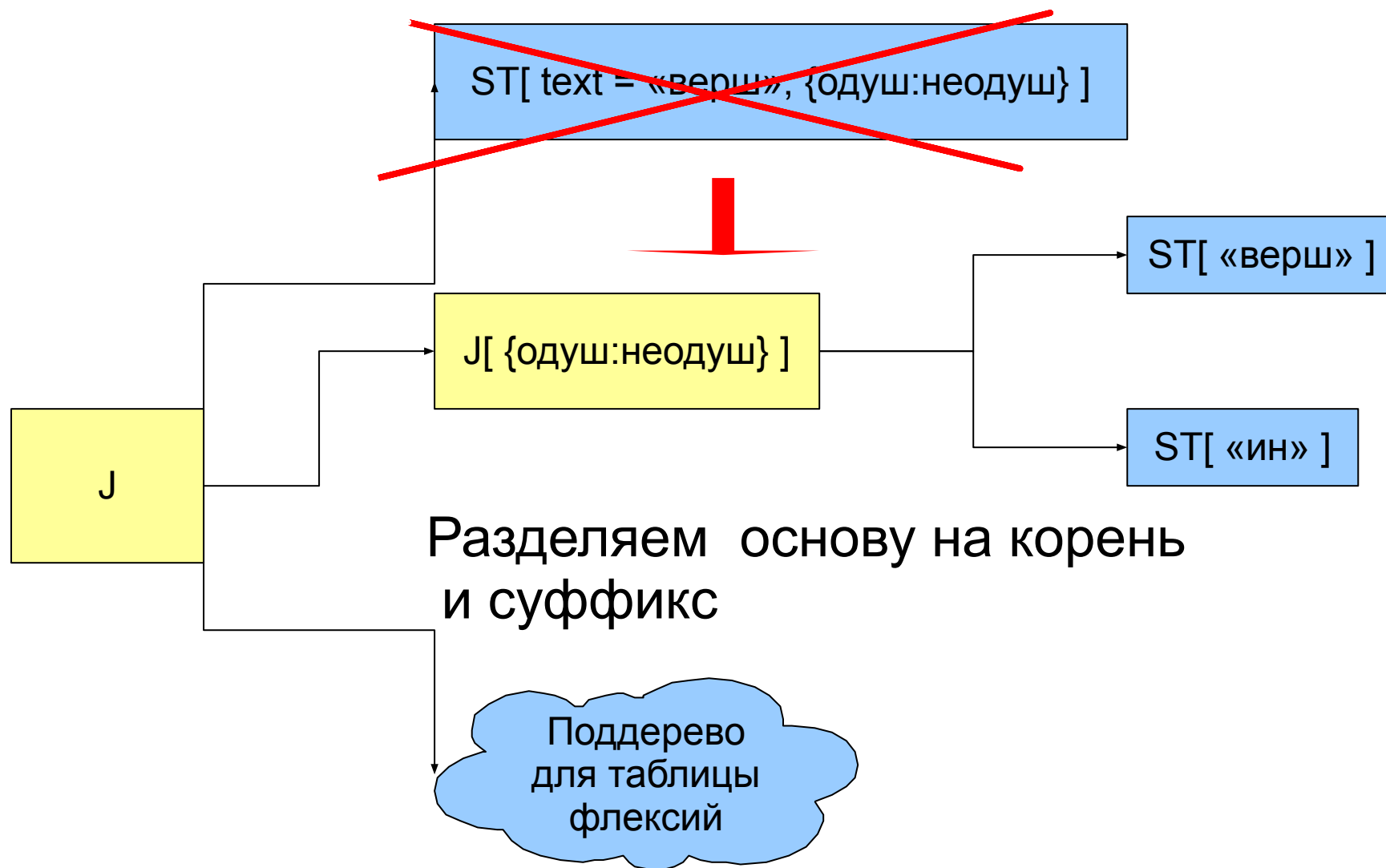
# Расширяемость словаря

- Предложенный подход позволяет легко расширять словарь
  - «В ширину» - добавление новых данных в существующей схеме. **Добавление новых основ и флексий.**
  - «В глубину» - добавление качественно новой информации. **Добавление семантической информации.**

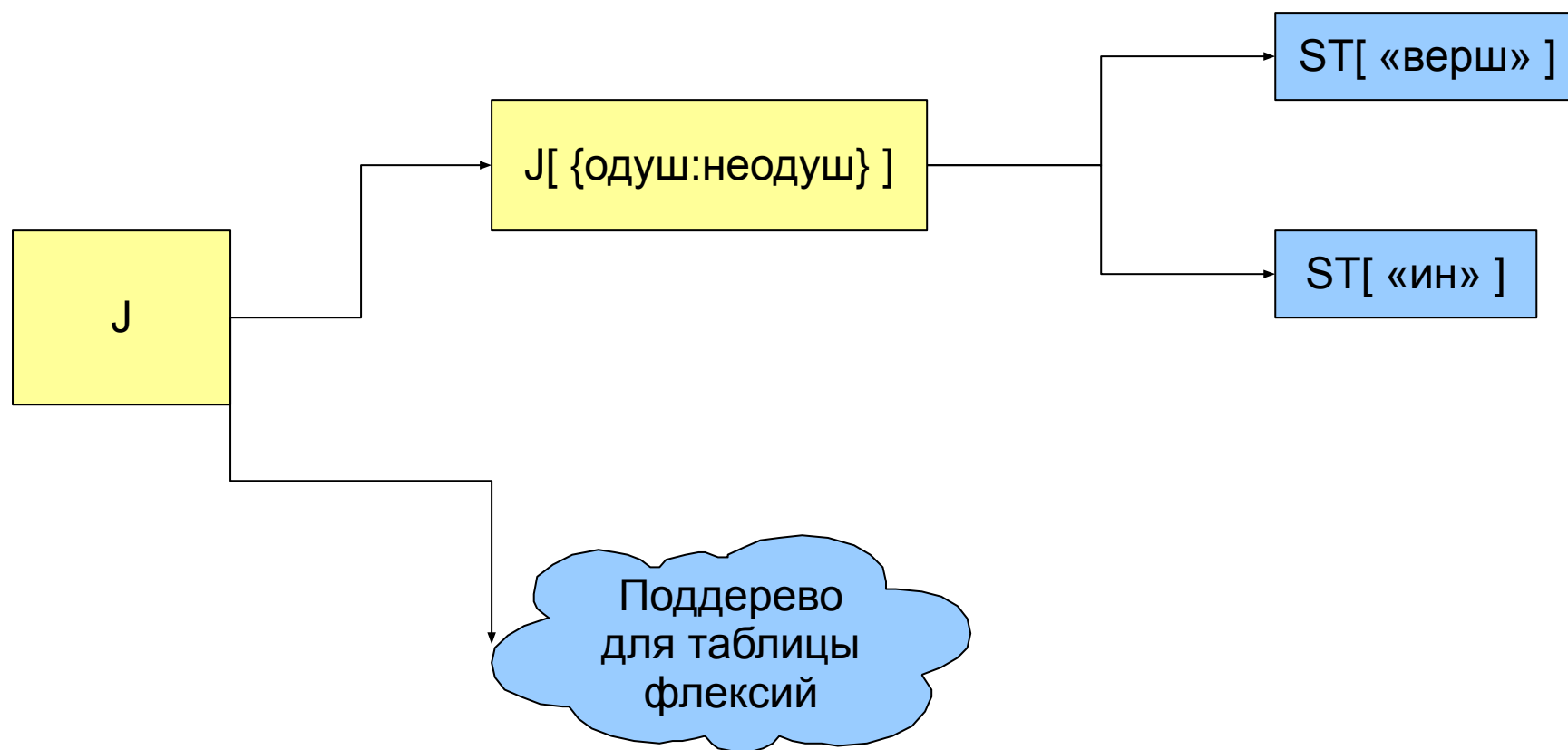
# Расширяемость словаря



# Расширяемость словаря

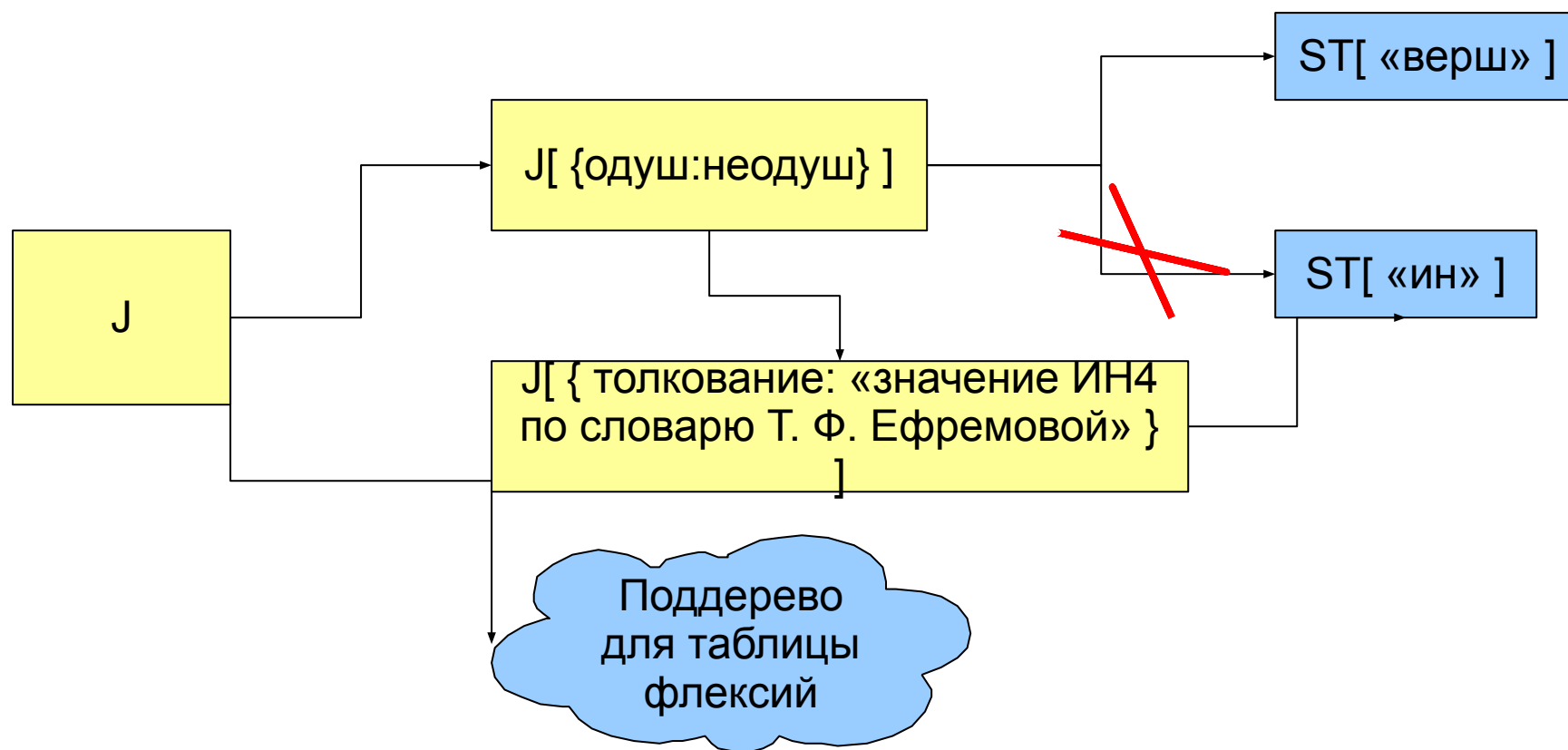


# Расширяемость словаря

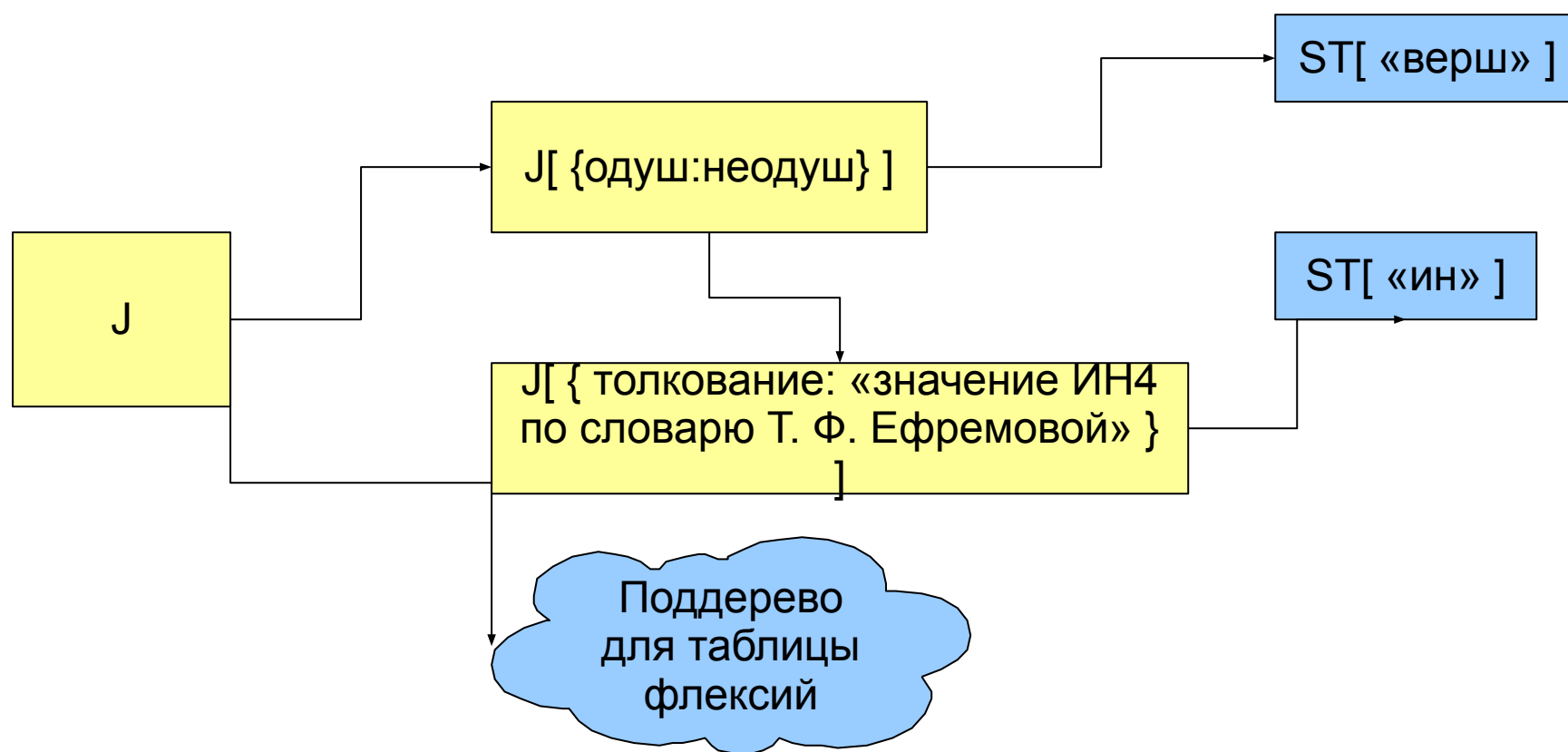


# Расширяемость словаря

Добавляем семантическую информацию для суффикса



# Расширяемость словаря





# Реализация словаря

- Модель реализована на основе некоторой объектной библиотеки
- Каждый R-объект является объектом в смысле программном смысле, он инкапсулирует:
  - Свойства R-объектов
  - Методы запроса множества текст-объектов, возможно, с заданными ограничениями

# Плюсы и минусы подхода

- Достаточно простой и мощный подход
- Унифицированное представление для различных задач
- Расширяемость «в ширину» и «в глубину»
- Возможность использования как модели для анализа
- Высокая вычислительная сложность при запросе элементов узла
- Кое-где модель неоправданно усложнена
- Опасность роста сложности модели при росте ее объема

