Лингвистическая база знаний для системы автоматизации построения онтологий

Мельничук Сергей
ДА-61

Автоматизация построения онтологий

- **Задача:** автоматизация построения онтологий из текстов на естественном языке
- **Решение:** методы обработки текстов на естественном языке

- **Необходимо:** лингвистическая информация (знания) о языке обрабатываемого текста

Автоматизация построения онтологий

Лингвистическая БЗ. Актуальность

Частичное формализированное описание ЕЯ, рассмотрение метода расширения и дополнения данного описания.

Актуально в рамках общей задачи автоматизации построения онтологий.

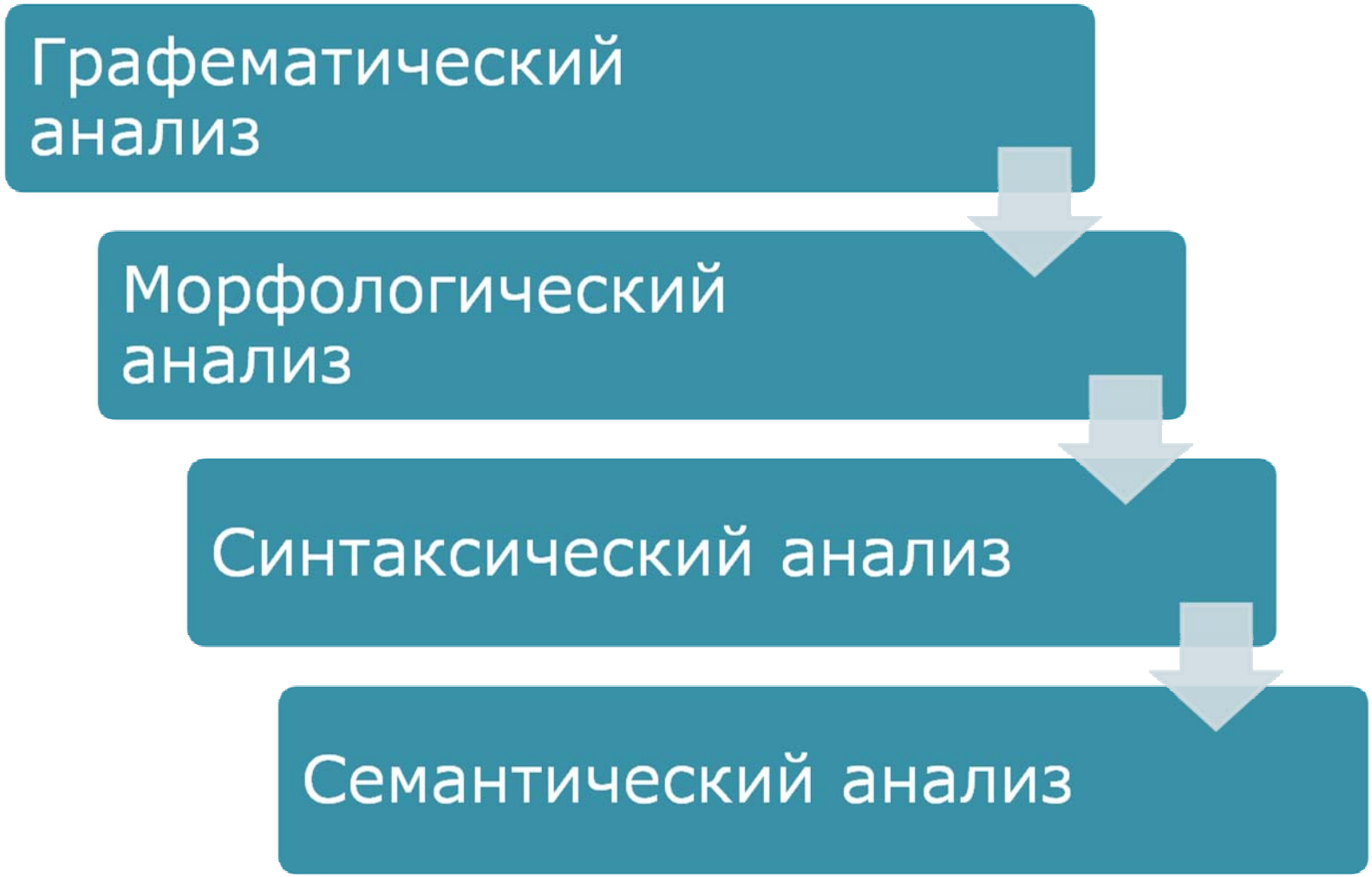
Лингвистическая БЗ. Задачи и цели

- **Цели:**
  - Описать лингвистическую информацию ЕЯ в терминах БЗ (сущности, связи, правила вывода информации)
  - Рассмотреть подход к дополнению и расширению БЗ

Лингвистическая БЗ. Задачи и цели

- **Задачи:**
  - Для каждого из этапов анализа текста на ЕЯ провести обзор необходимых для его реализации источников данных и знаний о ЕЯ.
  - Провести обзор полученных источников и рассмотреть возможности моделирования информации, которая хранится в указанных источниках.
  - Рассмотреть подход к описанию извлеченной из источников информации в терминах лингвистической базы знаний. Описать связи между сущностями БЗ и методы пополнения БЗ.

Анализ текста на ЕЯ



Графематический анализ

- Разделение входного текста на слова, разделители и т.д.
- Сборка слов, написанных в разрядку;
- Выделение устойчивых оборотов, не имеющих словоизменительных вариантов;
- Выделение ФИО (фамилия, имя, отчество), когда имя и отчество написаны инициалами;
- Выделение электронных адресов и имен файлов;
- Выделение предложений из входного текста;
- Выделение абзацев, заголовков, примечаний.

Графематический анализ

- Входные данные: plain-text (Windows-1251)
- Выходные данные: (Предложение «Иван спал.»)

Часть входного текста	Графематические дескрипторы
Иван	RLE Aa NAM?
спал.	RLE aa SENT_END

Графематический анализ

Графематические дескрипторы:

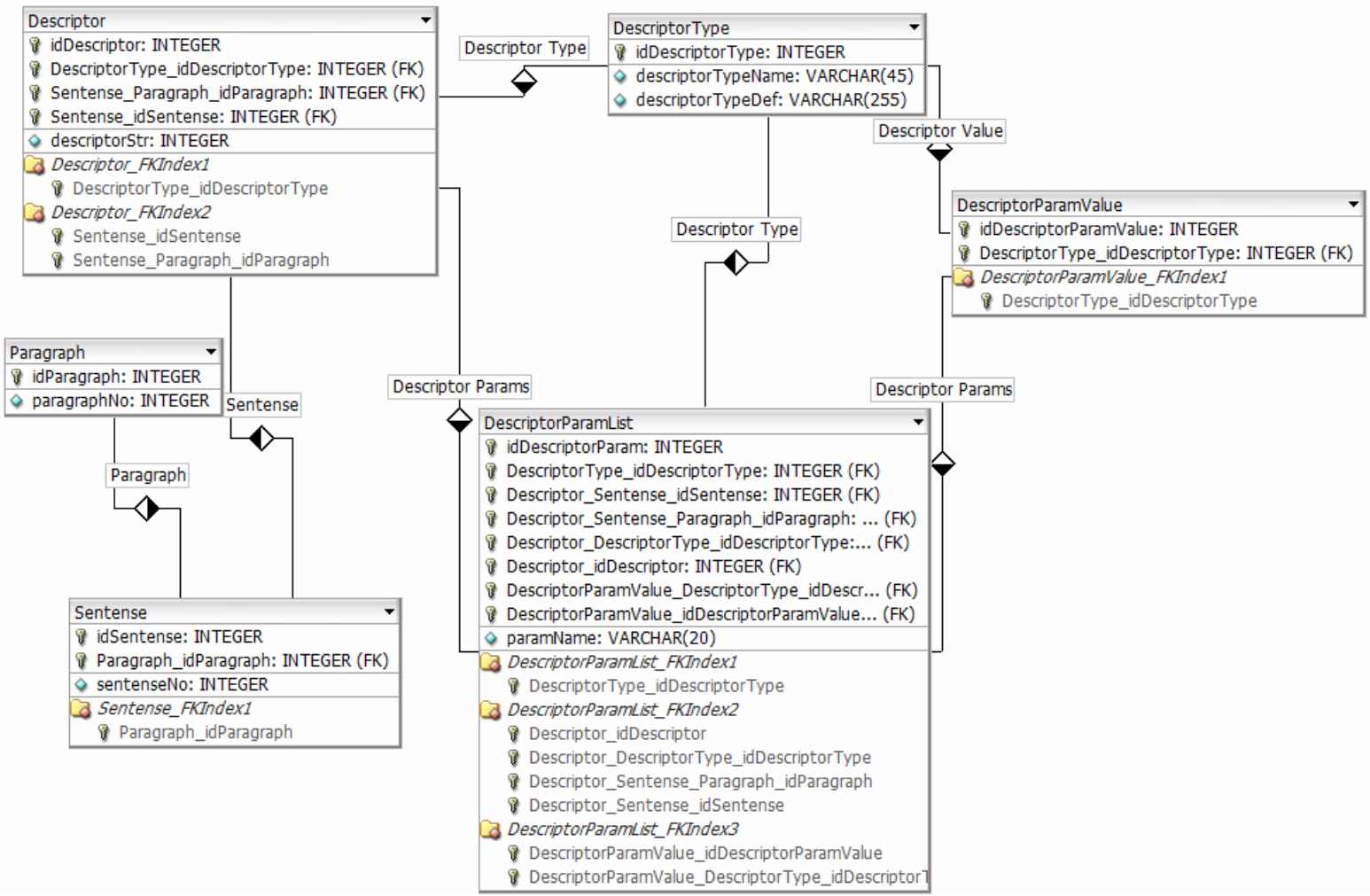
- RLE - русская лексема, присваивается последовательностям, состоящим из кириллицы ('*Иван*')
- DEL – разделитель ('*', '=', '_')
- DC - цифровой комплекс, присваивается последовательностям, состоящим из цифр ('1234')
- DSC - цифро-буквенный комплекс, присваивается последовательностям, состоящим из цифр и букв ('34h')
- И т.д.

Графематический анализ

Контекстные дескрипторы:

- NAM? - признак того, что лексема, возможно, является частью имени собственного. Присваивается лексеме, начинающейся с большой буквы и не имеющей перед собой символа конца предложения
- INDENT - ставится на начале абзаца
- FAM1/ FAM2 – начало/конец ФИО ("Иванов И.И.")
- FILE1/ FILE2 – начало/конец имени файла ("c:\test.txt")
- ABB1/ ABB2 – начало/конец аббревиатуры ("и т.п.")
- EA – электронный адрес ("www.aot.ru")
- И т.д.

Графематический анализ



Графематический анализ

Хранимые процедуры:

- add_descriptor_type – добавление нового типа дескриптора.
- add_descriptor_param – добавление нового параметра типу дескриптора.
- list_descriptors – вывод списка дескрипторов с описаниями и списками параметров.

Графематический анализ

Пример запроса

```
SELECT
    d1.descriptorStr,
    d2.descriptorTypeName
FROM
    Descriptor d1,
    DescriptorType d2,
    DescriptorParamList d3,
    DescriptorParamValue d4
WHERE
    d1.DescriptorType_idDescriptorType = d2.idDescriptorType
    AND d2.descriptorTypeName = “RLE”
    AND d1.idDescriptor = d3.Descriptor_idDescriptor
    AND d4.idDescriptorParamValue =    d3.DescriptorParamValue_id DescriptorParamValue
    AND d4.idDescriptorParamValue = “NAM?”
```

Результат запроса:

descriptorStr	descriptorTypeName
Иван	RLE

Морфологический анализ

На основе входной словоформы:

- Получение леммы и ее свойств
- Получение морфологических характеристик входной словоформы

Морфологический анализ

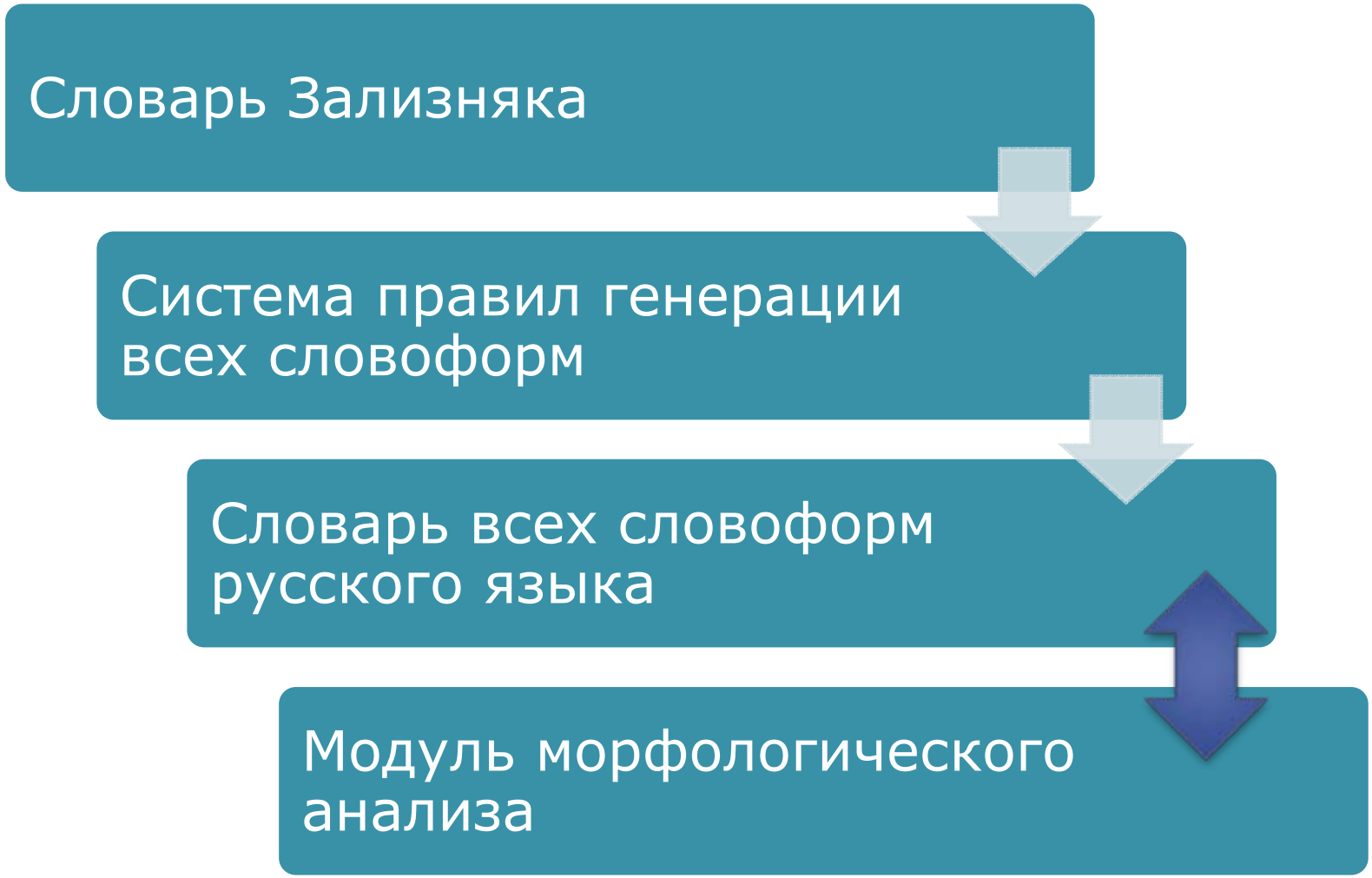
- Входные данные:
  Словоформа (слово)

- Выходные данные:
  - псевдооснова слова (общая для всех словоформ данного слова подстрока)
  - Морфологические характеристики словоформы (род, падеж, часть речи, число, и т.д.)

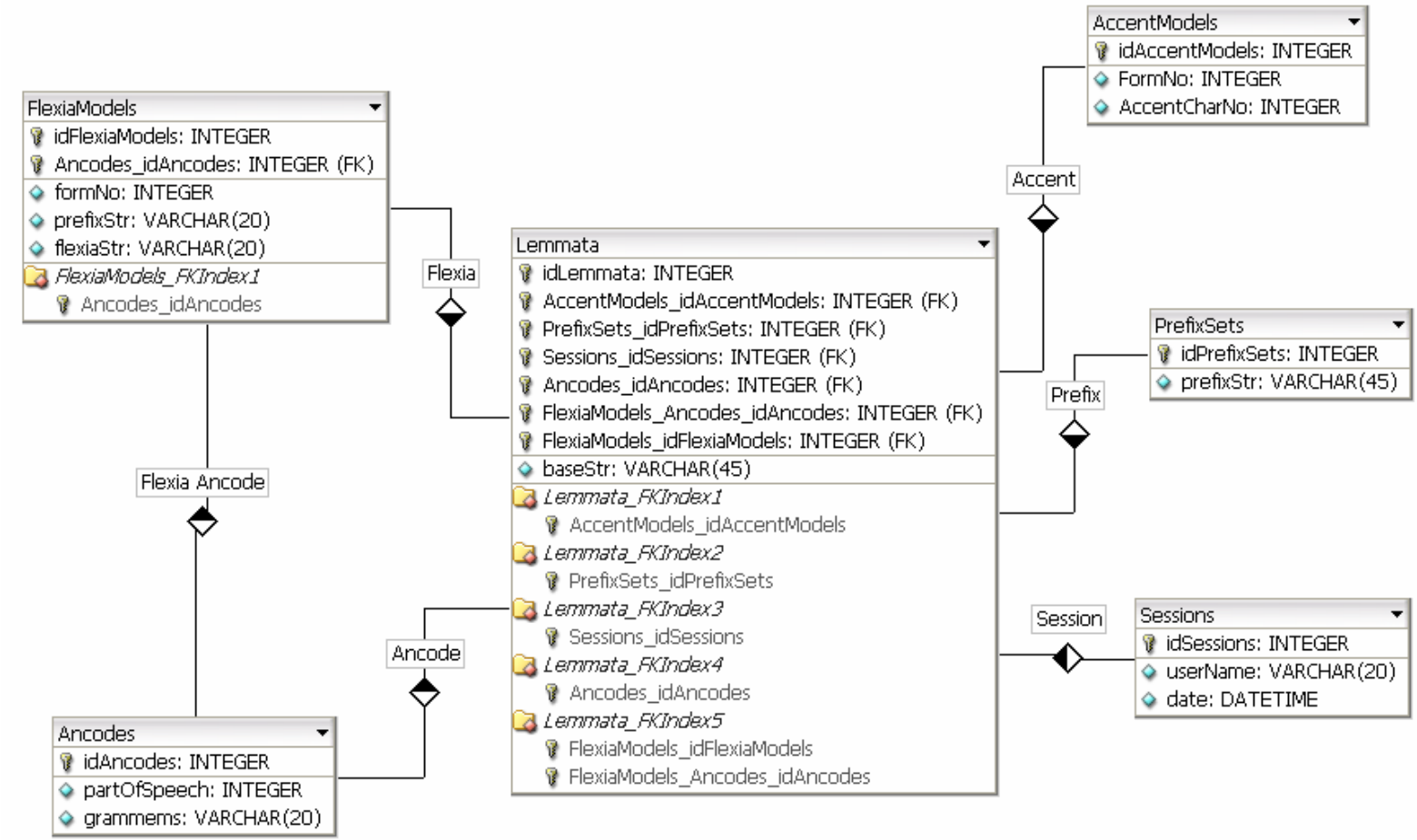
ра’нний П мр,ед,им,од,но,

по|ра’ньше П од,но,сравн,

Морфологический анализ



Морфологический анализ



Морфологический анализ

Пример запроса всего слова по индексу префикса (Prefix_Number) и индексу словоформы (Form_Number):

```
SELECT
    P.PrefixStr,
    F.PrefixStr,
    L.BaseStr,
    F.FlexiaStr,
FROM
    Lemmata L,
    FlexiaModels F,
    PrefixSets P,
WHERE
    L.idFlexiaModel = F.idFlexiaModel
    AND F.FormNo = Form_Number
    AND P.idPrefixSet = Prefix_Number
```

Результат запроса:

P.PrefixStr	F.PrefixStr	L.BaseStr	F.FlexiaStr
недо	пере	определ	или

Морфологический анализ

Хранимые процедуры:

- **find_by_lemma** – вывод всех слов и их характеристик, найденных в словаре с указанной основой слова
- **add_lemma, add_ancode, add_accent_model, add_prefix_set** – добавление новой леммы, грамеммы, ударения, префикса

Синтаксический анализ

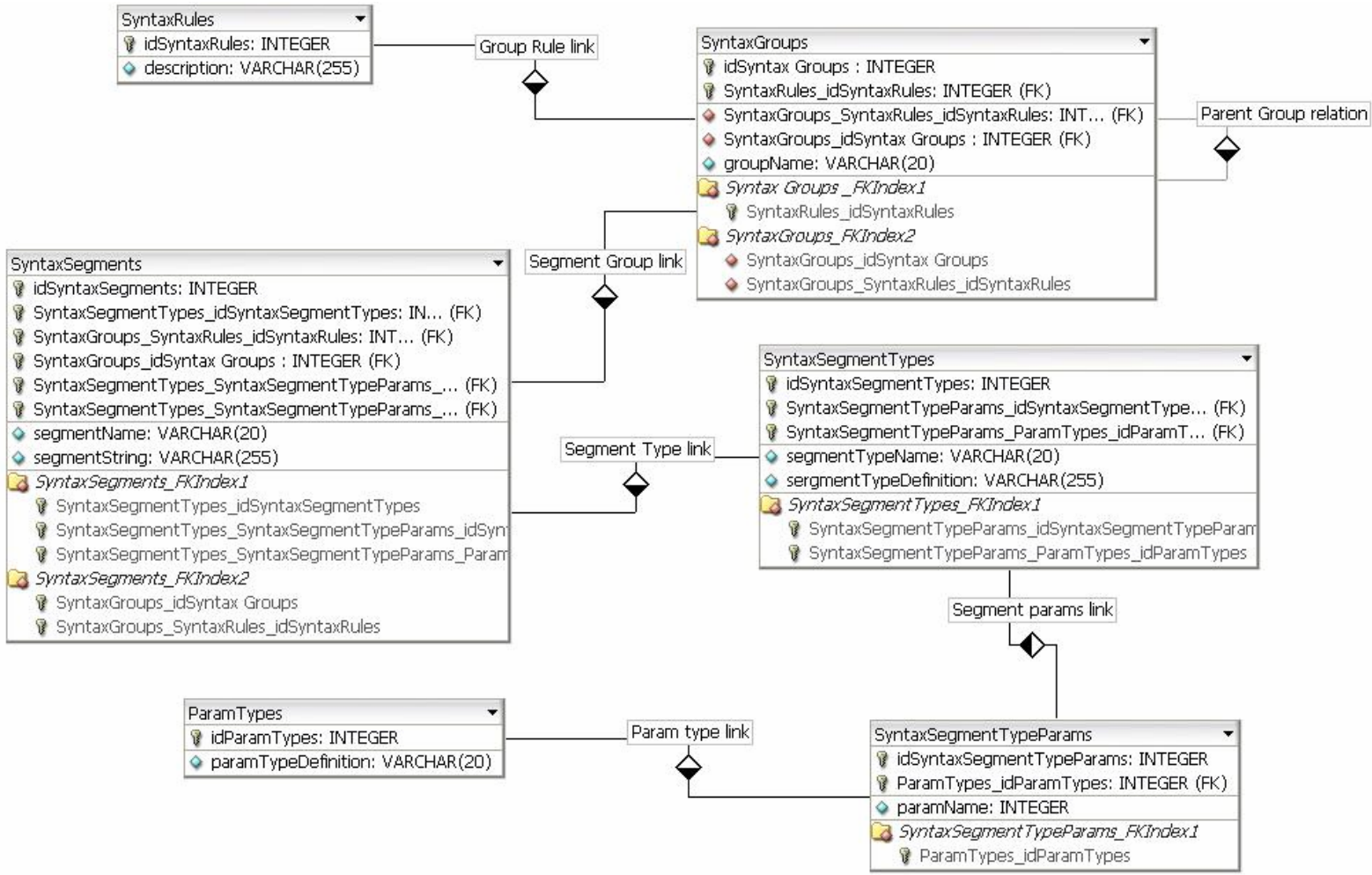
- Деление текста на синтаксические сегменты
- Определение синтаксических правил для каждого сегмента (Построение аналитических форм глагола внутри исходных сегментов)
- Построение иерархии сегментов (создание групп сегментов и связей между ними)

Синтаксический анализ

Типы синтаксических правил:

- КОЛИЧ - Количественная группа (последовательность числительных): «Двадцать восемь»
- ПГ - Предложная группа: «В дом, на холме»
- ПРИЛ-СУЩ - Группа существительного, пре-модифицированная одним или несколькими прилагательными: «Длинная тяжелая дорога, двигающийся человек»
- СУЩ-ЧИСЛ - Группа существительного, пре-модифицированная числительным: «Восемь попугаев, два человека»
- НАРЕЧ_ГЛАГОЛ - Глагол, пре-модифицированный наречием «злостно нарушает, тяжело жить»
- НСО - необособленного согласованного определения.
- И т.д.

Синтаксический анализ

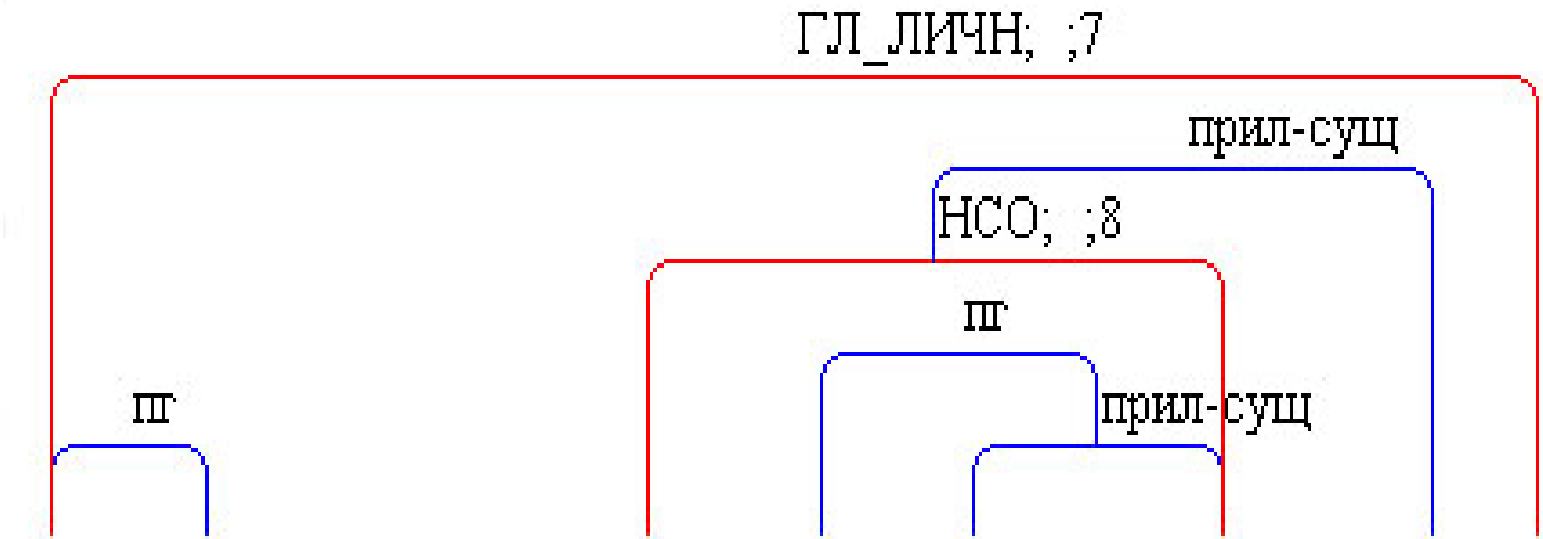


Синтаксический анализ

Хранимые процедуры:

- **add_group** – добавление новой синтаксической группы.
- **add_segment_type** – добавление нового синтаксического сегмента.
- **add_rule** – добавление нового синтаксического правила.

Синтаксический анализ



Семантический анализ

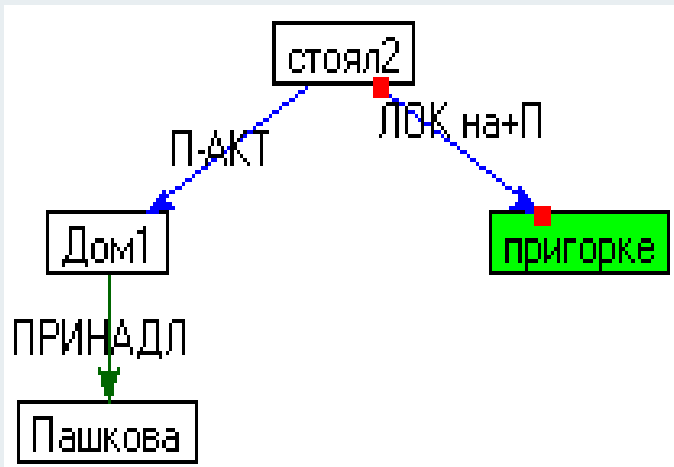
- Определение семантических узлов
- Определение семантических атрибутов узлов
- Определение семантических связей между узлами

Семантический анализ

- Входные данные:

  Результаты морфологического и синтаксического анализа предложения («Дом Пашкова стоял на пригорке»).

- ```
  ПРИНАДЛ (Пашков, дом),
  П-АКТ(дом, стоял),
  ЛОК (пригорке, стоял).
  ```

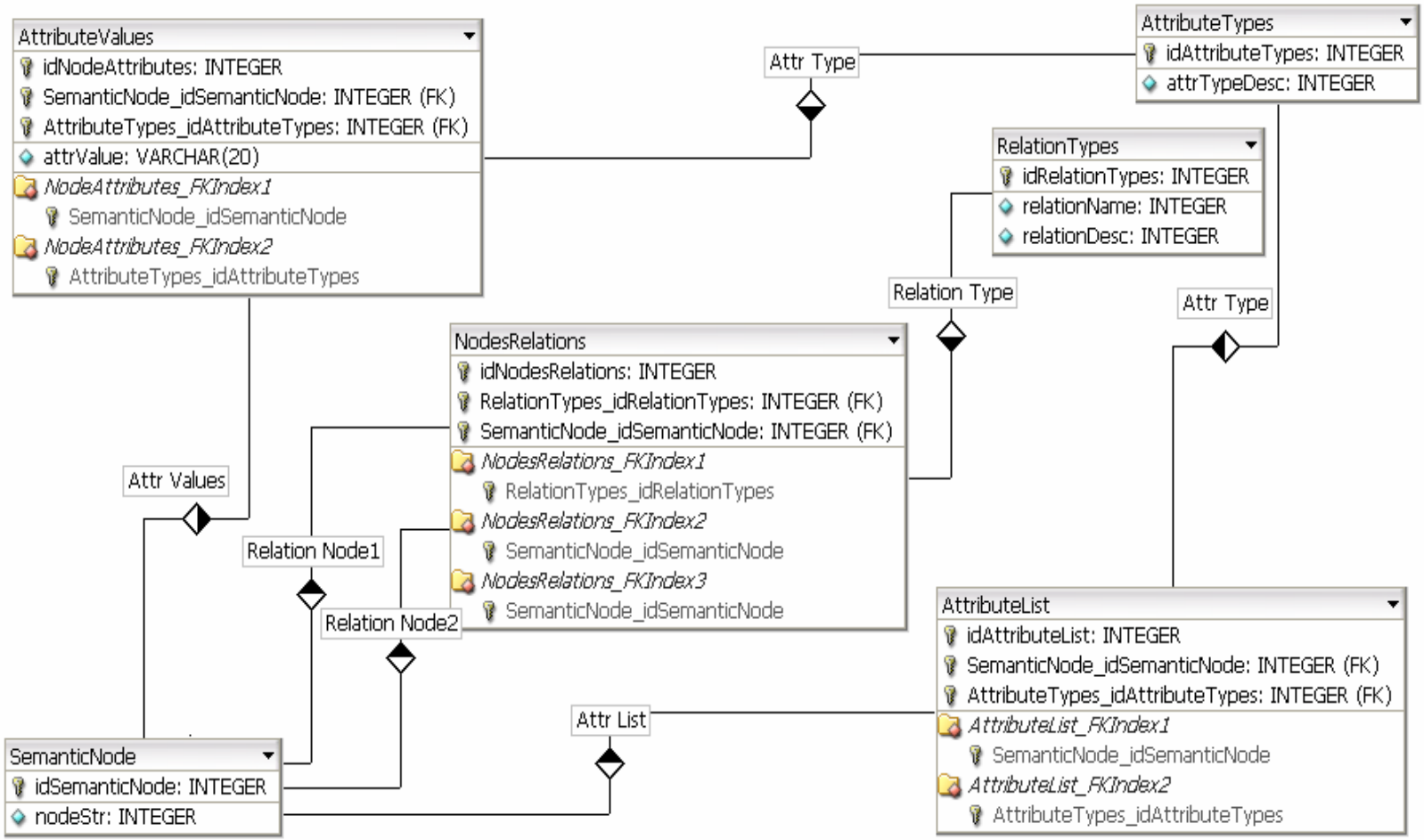

  ```

Семантический анализ

Классы семантических отношений:

- родо-видовые отношения;
- отношения «целое — часть»;
- синонимия и антонимия;
- логические отношения;
- функциональные отношения;
- атрибутивные отношения;
- количественные отношения;
- пространственные отношения;
- временные отношения;
- лингвистические отношения.

Семантический анализ



Семантический анализ

Хранимые процедуры:

- **add_relation_type** – добавление нового типа семантической связи.
- **add_attribute** – добавление нового атрибута для семантического узла.
- **add_attribute_type** – добавление нового типа атрибута.

Семантический анализ

Пример запроса для вывода списка синонимов

```
SELECT
   N2.nodeStr
FROM
   SemanticNode N1,
   SemanticNode N2,
   NodesRelations R,
   RelationTypes T
WHERE
   T.relationName = "synonym",
   T.idRelationType = R.idRelationTypes,
   R.idSemanticNode1 = N2.idSemanticNode,
   R.idSemanticNode2 = N1.idSemanticNode,
   N1.nodeStr = Input_Str
```

Заключение

- Представлены источники знаний/данных, необходимых на каждом этапе анализа на ЕЯ.
- Представлены схемы БЗ для каждого этапа анализа текста на ЕЯ.
- Представлен подход к дополнению БЗ новыми данными/знаниями.

Спасибо за внимание