УДК 004.822:004.89

# Методика разработки лексико-семантических паттернов для извлечения терминологии научной предметной области

*Кононенко И.С. (Институт систем информатики СО РАН),*
*Сидорова Е.А. (Институт систем информатики СО РАН)*

В статье описывается подход к автоматизации извлечения терминологии для пополнения онтологии научной предметной области из текстов на русском языке. Применимость методов автоматического пополнения онтологии из текстов на естественном языке зависит от характеристик корпуса текстов и используемого языка. Специфика входного языка, характеризующегося сильной флективностью и свободным порядком слов, и отсутствие большого корпуса текстов приводят к выбору лингвистического подхода, базирующегося на использовании лексико-семантических паттернов. К особенностям предлагаемой методики извлечения информации относятся а) автоматическое пополнение предметного словаря на основе онтологии и корпуса текстов и разметка его с помощью системы семантических признаков; б) определение небольшого набора исходных структурных мета-паттернов, устанавливающих концептуальные контексты извлечения онтологической информации; в) автоматическое порождение по набору структурных мета-паттернов множества лексико-семантических паттернов, определяющих лексические, семантические и синтаксические свойства контекстов извлечения.

***Ключевые слова:*** *построение онтологии, лексико-семантический паттерн, извлечение терминов, генерация паттернов.*

## 1. Введение

В настоящее время для формализации и систематизации знаний и данных в научных предметных областях (НПО) активно используются онтологии, которые позволяют описать научную дисциплину или область научных знаний во всех ее аспектах, включая характерные объекты и предметы исследования, применяемые научные методы, выполняемые проекты и полученные результаты [3]. Процесс разработки такой онтологии состоит из нескольких этапов, главными из которых являются построение терминологической части онтологии,

предполагающее создание таксономий понятий и отношений и описание их свойств, и пополнение онтологии, т.е. добавление в нее экземпляров понятий и отношений. Если первый этап задает скелет онтологии, то второй этап наполняет ее содержанием.

Чтобы получить онтологию, которая бы достаточно полно описывала НПО, требуется обработать огромное количество научных публикаций и информационных ресурсов, содержащих сведения из моделируемой области. Для облегчения и ускорения этого процесса разрабатываются методы автоматического пополнения онтологии на основе текстов на естественном языке [9, 13] и web-документов [6, 7]. Для автоматической обработки текстов используются подходы на основе кластеризации, в которых применяются широко известные кластерные и статистические методы, и подходы на основе шаблонов, в которых используются лингвистические шаблоны. Однако первый подход для хорошей работы требует наличия больших корпусов текстов, поэтому более распространены методы на основе лингвистических шаблонов.

У истоков лингвистического подхода стоит предложенная в работе [10] идея о возможности автоматизации построения семантических связей на основе диагностических контекстов, представленных в виде лексико-синтаксических шаблонов. Данный метод, известный как шаблоны Херст (Hearst patterns), предназначен для обработки неструктурированных англоязычных текстов. Он широко применялся для извлечения родовидовых отношений и предполагал извлечение из коллекции документов упорядоченных пар слов, соответствующих множеству заранее составленных шаблонов. Подход M. Hearst использовался и совершенствовался многими другими исследователями, а также применялся для других языков.

В ряде работ предлагается формальный аппарат для записи лексических и лексико-синтаксических шаблонов. Так, в работе [5] формулируется XML-схема языка для формализации лексико-синтаксических шаблонов, используемых для пополнения онтологий. Система Alex [2] и являющийся ее развитием инструмент DigLex [4] предоставляют довольно гибкие средства описания слов и словосочетаний в виде шаблонов, которые используются затем для автоматического распознавания этих единиц в тексте. Они расширяют возможности традиционных лексикографических систем: язык описания шаблонов поддерживает использование альтернатив, ссылки на шаблоны, повторители, условия на контекст, дистантный контекст и т.п. Язык позволяет записывать правила не только для распознавания текстовых объектов, но и для определения их лексических и семантических атрибутов. Однако в языках Alex и DigLex нет встроенных средств для указания грамматических признаков распознаваемых лексических единиц и грамматического

согласования нескольких единиц, необходимых для однозначного выделения языковых конструкций (например, именных групп). Этого последнего недостатка лишен предложенный в работе [1] язык LSPL, позволяющий задавать грамматические свойства входящих в него элементов.

Исследования, решающие задачу автоматического или полуавтоматического пополнения онтологии, опираются на паттерны (шаблоны), которые отображают языковые структуры, встречающиеся в текстах, в соответствующие элементы онтологии (понятия, отношения, экземпляры понятий и отношений). Это лексико-синтаксические паттерны, которые используют лексические представления и синтаксическую информацию [12, 14] или лексико-семантические паттерны, которые в процессе извлечения объединяют лексические представления с синтаксической и семантической информацией [11, 15].

В статье описывается подход к автоматизации пополнения онтологий НПО, базирующийся на использовании лексико-семантических паттернов (ЛСП) как расширения лексико-синтаксических паттернов онтологического проектирования. Особенностью данного подхода является то, что применяемые в нем ЛСП автоматически строятся на основе паттернов онтологического проектирования (паттернов ОП) других типов [8], входящих в систему автоматизации разработки онтологий на основе разнородных паттернов онтологического проектирования [16]. Еще одной отличительной чертой описываемого подхода является ориентация на русский язык, который, как и многие другие славянские языки, является сильно флективным языком со свободным порядком слов, что предполагает акцент на вариативности способов представления элементов онтологии в тексте и поиск методик генерации альтернативных средств выражения для максимизации полноты извлечения.

## 2. Особенности предметного словаря для пополнения онтологий

Для автоматического пополнения онтологии с помощью ЛСП требуется обеспечить извлечение из текста специфических терминов данной НПО. При этом для извлечения заранее известных терминов используются словари универсальной и предметной лексики, а для извлечения новых терминов (в частности, наименований объектов НПО или специфичных предикатных слов) – специализированные терминологические паттерны (Т-ЛСП) .

Предметный словарь – это объем лексики, организованной по семантическому принципу с отражением определенного набора базовых формальных отношений (см. Рис. 1). В словарной статье хранится вся необходимая информация как для извлечения термина из

текста, так и для поддержки последующих этапов анализа текста. Каждый термин, найденный в тексте с помощью предметного словаря, снабжается морфологической и семантической информацией, которая в дальнейшем используется в процессе применения ЛСП.

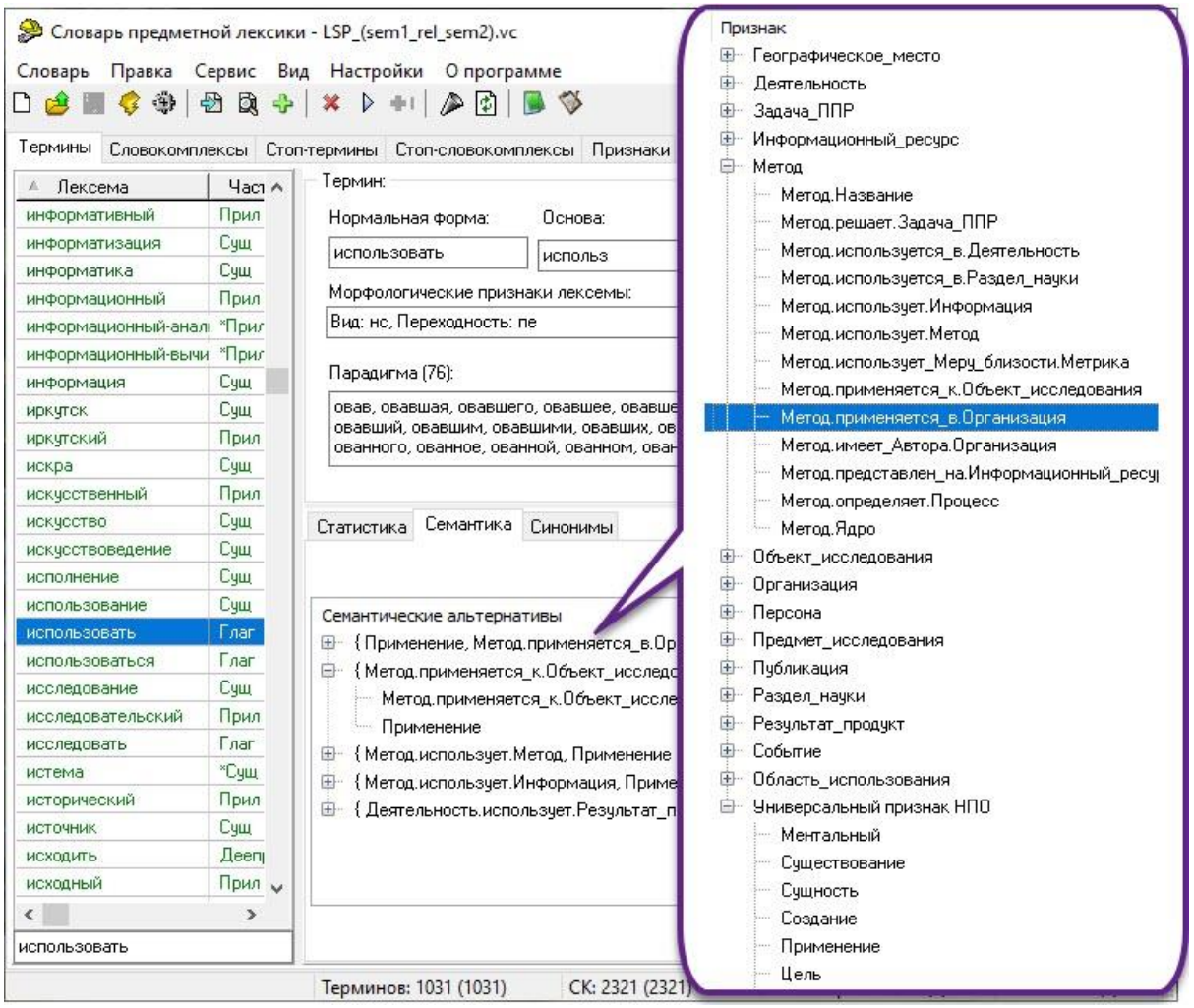


*Рис. 1. Предметный словарь.*

Снабжение терминов ПО лексико-семантическими характеристиками опирается на универсальную иерархию классов терминов, построенную вручную для научной области в целом, и иерархию предметных лексико-семантических классов, генерируемую автоматически на основе структуры заданной онтологии.

Для автоматической генерации словаря создана методика автоматического формирования системы лексико-семантических характеристик на основе имен онтологии. Предметный словарь создается как расширение словаря общенаучной лексики и включает две независимые иерархии лексико-семантических классов: универсальная иерархия

(*Универсальный признак НПО*) и предметно-ориентированная иерархия, создаваемая на основе онтологии НПО.

## 2.1. Универсальная иерархия признаков словаря

Универсальная иерархия включает признаки для разметки следующих групп терминов:

a) названия основных сущностей научных текстов, такие как *ученый, проект, организация, научный результат*;

b) названия типовых объектов (например, разделов науки – *информатика, физика*);

c) предикаты, используемые для описания (*Существование*) или связывания (*Применение*) сущностей; предикаты представлены, как правило, глаголами, включая личные, причастные и деепричастные формы, глагольными группами с зависимым инфинитивом (*использовать - решить использовать*), глаголами с зависимым предлогом (*применять к*), реализуемым как N-грамма, субстантивированными отглагольными именами (*использование*), существительными (*название*), ролевой лексикой (*разработчик, автор*), устойчивыми лексическими оборотами, реализуемыми как N-грамма (*под названием)*;

d) вспомогательная лексика (*Вспом*), включая связочные (*быть*, *являться*), фазовые (*начать*), модальные (*решить*, *позволять*) глаголы, семантически пустые глаголы типа лексических функций (*получать*, *давать*), а также слова и лексические конструкции, необходимые для представления универсальных категорий, таких как фамилии и имена людей, дат, типовых сокращений (*км, д.н., NLP, ПО, ОАО*) и т.п.

## 2.2. Генерация лексико-семантической иерархии словаря

Предметно-ориентированная иерархия лексико-семантических классов включает онтологически обусловленные признаки, названия которых формируются на основе следующих наименований.

- Имена классов паттерна(ов) содержания и соответствующие имена классов наследников, представленных в рассматриваемой онтологии НПО:

  **имя_класса**

  для терминов, являющихся именами классов, например: *Математический метод, Нормирование, Метод визуализации*;

  **имя_класса.Ядро**

  для слов, являющихся вершинами имен классов, например, слово *метод* снабжается признаком *Метод исследования.Ядро*

- Имена признаков-атрибутов класса, формируемых по шаблонам:

**имя_класса.имя_признака**

для значений атрибутов, например: *Метод.Название, Персона.Фамилия, Организация.Дата_основания*;

**имя_класса.имя_признака.type**

для названий самих атрибутов, например: *Метод.Название.type, Метод.Описание.type, Организация.Дата_основания.type*;

- Имена признаков-отношений между классами, формируемых по шаблону:

  **имя_класса.имя_отношения.имя_класса**

  например: *Метод.применяется_к.Объект_исследования,*

  *Организация.расположена_в.Географическое_место*

Все термины словаря универсальной научной лексики размечены признаками из универсальной иерархии. Новые термины из конкретной научной ПО размечаются с помощью признаков из предметно-ориентированной иерархии.

Термины конкретной НПО, присутствующие в словаре универсальной научной лексики, получают синкретический признак с одновременно выраженным значением универсального и предметного лексико-семантического класса. Так, для глагола ‘*использовать*’ выделено пять синкретических признаков (см. Рис. 1), каждый из которых включает универсальный класс *Применение* в сочетании с онтологически обусловленным признаком:

*Метод.применяется_в.Организация,*

*Метод.применяется_к.Объект_исследования,*

*Метод.использует.Метод,*

*Метод.использует.Информация,*

*Деятельность.использует.Результат_продукт*

Лексико-семантические признаки используются при описании ЛСП (в аргументах и результате) как способ обращения к терминам ПО с определенной семантикой.

## 3. Терминологические лексико-семантические паттерны

Для пополнения онтологии для каждого ее класса, представляющего понятие моделируемой предметной области, строится набор лексико-семантических паттернов, описывающих различные способы представления соответствующей ему информации в научных текстах. В данной работе рассматриваются терминологические паттерны (Т-ЛСП), которые используются для извлечения из текстов новых терминов ПО, не заданных в словаре.

Каждый ЛСП реализует модель вида: <Arguments, Constraints, Results>, где Arguments – множество семантических аргументов факта, которым сопоставляются термины ПО, Constraints – семантические, синтаксические и/или позиционные условия на аргументы, а Results описывает результат применения ЛСП, которым может быть новый термин и его семантических класс. Таким образом, Т-ЛСП представляют собой лексико-семантические шаблоны, формируемые на основе опорных терминов/маркеров с указанием семантических и грамматических ограничений.

Для извлечения новых терминов предложены два типа паттернов.

1) Паттерны первого типа моделируют именную группу (ИГ), в вершине которой представлено существительное или именная группа – представитель класса онтологии, и позволяют извлекать имена индивидов, относящихся к данному классу.
2) Паттерны второго типа позволяют извлекать новые термины на основе контекста, в котором присутствуют индикаторы отношений (атрибутов), в качестве которых, как правило, выступают предикатные термины, сопоставленные названиям отношений или атрибутов класса онтологии. Т-ЛСП первой подгруппы данного типа представляют контексты, структурными составляющими которых являются смысловые компоненты: Субъект, Объект и Предикат. Т-ЛСП второй подгруппы данного типа представляют контексты, структурными составляющими которых являются смысловые компоненты: Объект, Атрибут, Значение.

Для того, чтобы обеспечить автоматическую генерацию паттернов разработан язык и предложена методика создания типовых паттернов (или мета-паттернов), в состав которых включаются переменные. Создание исходных структурных мета-паттернов для извлечения информации для конкретной онтологии осуществляется путем означивания переменных именами онтологических классов, атрибутов и отношений. Итоговые мета-паттерны, помимо семантической информации, содержат лексические маркеры, синтаксические и позиционно-структурные ограничения.

## 3.1. Извлечение наименований объектов на основе именных групп

Паттерны первого типа позволяют извлекать имена индивидов на основе центрального слова или термина с привлечением синтаксических правил сборки именных групп. Во множестве терминологических наименований наиболее частотны паттерны следующего вида (типовой пример):

[<Прил>*, X, [<Прил, рд>*, <Сущ, рд>]* ] ⇒ X.Название (1)

Данный Т-ЛСП включает три аргумента: 1) цепочку прилагательных <Прил>* (значок * в рамках языка паттернов означает цепочку компонент произвольной длины включая нулевую), 2) термин, имеющий лексико-семантический класс X (здесь и далее под X понимается не только имя класса, но и центральное слово, отмечаемое лексико-семантическим признаком *имя_класса.Ядро*) и 3) именную группу в родительном падеже, собранную по вложенному подшаблону вида [<Прил, рд>*, <Сущ, рд>]*; ограничениями здесь являются: a) указание семантического класса для 2-го аргумента, b) указание падежа для вложенного Т-ЛСП, используемого в качестве 3-го аргумента; результат Т-ЛСП определяет лексико-семантический признак *X.Название* для всех терминов, извлекаемых с помощью данного Т-ЛСП.

Подшаблон [<Прил, рд>*, <Сущ, рд>]* дает возможность учесть согласование *Прил* и *Сущ* по падежу для зависимой группы существительного в родительном падеже в рамках именных групп. Это позволяет избежать извлечения некорректных индивидов: **соответствующие этим проектам агрегированные оценки, *задачи принятия решений аксиоматические теории рационального поведения, *модели различные варианты.* К сожалению, язык шаблонов в его текущей версии не позволяет в общем случае учесть согласование существительного и зависимого прилагательного.

Данный типовой паттерн позволяет генерировать конкретные Т-ЛСП путем подстановки в качестве X имён классов онтологии. Например, для извлечения названия метода может быть автоматически сгенерирован следующий паттерн:

[<Прил>*, Метод, [<Прил, рд>*, <Сущ, рд>]* ] ⇒ Метод.Название

Этот паттерн позволяет извлечь такие термины: *метод опорных векторов, метод анкетного опроса, метод медиан рангов, метод нейронных сетей, метод анализа иерархий Саати, метод интервью, метод самооценки*.

## 3.2. Извлечение наименований объектов на основе индикаторов отношений

Вторая группа паттернов формируется на основе индикаторов отношений, в качестве которых, выступают термины из словаря, сопоставленные названиям отношений онтологии.

Исходные мета-шаблоны данного типа описываются в соответствии со следующим принципом. В структуре шаблона выделяются два известных компонента: упоминание отношения и одного из аргументов данного отношения (указываются семантические классы

терминов и их грамматические признаки), а также третий – неизвестный компонент, который и требуется извлечь (помечается переменной вида ***$t***).

[X.REF, X.Rel.Y, $t<ИГ>] ⇒ Y.Название (2)

и симметричное

[ $t<ИГ>, X.Rel.Y, Y.REF] ⇒ X.Название

Отношение Rel связывает объекты классов X и Y (в онтологии X.Rel - Object Property).

При генерации учитываются следующие особенности конструкций.

a) Третий компонент представлен именной группой, все ИГ извлекаются по шаблонам, аналогичным шаблонам из п.1, в которых в качестве X указано *Сущ*.

b) Для известного объекта отношения заводится служебный шаблон, объединяющий три варианта: имя класса, название и ядро имени класса:

[X] [ X.Название] [ X.Ядро ] => **X.REF**

c) В качестве имени отношения выступает глагольная группа (ГГ), которая может включать вспомогательный глагол в спрягаемой форме и значимый предикат в форме инфинитива:

ГГ = [<Глаг>] [Вспом<Глаг>, <Инф>]

d) В большинстве шаблонов данного типа будет использоваться разрыв (обозначаем в мета-паттернах ***gap***), ограничения для которого задаются следующим паттерном:

gap = [s/"."] [s/"!"] [s/"?"] [s/":"] [s/";"] [s/"("] [s/")"] [s/"/”"] [s/"/n"] [s/","]

Разработана типология шаблонов описываемой группы, которые строятся на базе приведенных исходных схем паттернов для представления всего разнообразия конструкций. ЛСП данной группы представлены в текстах преимущественно глагольными и – существенно реже – субстантивными конструкциями. Подробно рассмотрены паттерны с глагольными формами в рамках простого или сложного предложения.

Варианты возможных контекстов, описываемых паттернами, определяют следующие факторы: (1) синтаксические ограничения, т.е. грамматические классы и согласование грамматической информации, (2) позиционно-структурные ограничения, т.е. порядок следования компонент и принадлежность компонент одной/нескольким клаузам в рамках предложения.

В результирующих шаблонах учитываются различные варианты порядка слов и возможные разрывы между компонентами исходной типовой схемы. В качестве ограничений на элементы в разрывах используются отсутствие отрицаний, разделителей предложений и других знаков пунктуации.

### 3.2.1. Активные конструкции с переходным глаголом

Структурными составляющими трехчленной активной конструкции являются смысловые компоненты Субъект, Объект и Предикат. В активной конструкции:

- субъект личной формы глагола выражается именительным падежом
- объект выражается винительным падежом,
- предикат представлен глагольной формой, соответствующей действительному залогу.

Учитывая бинарность отношения, извлекаемая сущность может быть указана а) как переменная в позиции субъекта, если объект представлен явно (известным термином) или б) как переменная в позиции объекта, если известен термин в позиции субъекта. Такие симметричные конструкции имеются для каждого типа глагольных форм. При этом разрыв допускается только между известными (явно заданными) компонентами контекста. Ниже представлены варианты метапатернов для простого предложения (a-d, g-i) и сложного предложения с придаточным определительным (d-e). Предикат простого предложения или придаточного определительного выражен переходным личным глаголом или зависимым от вспомогательного слова инфинитивом в составе группы сказуемого:

ГГ1 = [<Глаг, пе>] [<Глаг, Вспом>, <Инф, пе>]

Кроме того, шаблоны покрывают причастные и деепричастные обороты в рамках простого предложения.

1) Личная форма глагола и инфинитив

a. [X.REF<им>, gap, X.Rel.Y<ГГ1>, $t<ИГ, вн>] ⇒ Y.Название

*// методы моделирования выполняют сбор необходимых научных данных и их систематизацию*

Симметричная конструкция:

[ $t<ИГ,им>, X.Rel.Y<ГГ1>, gap, X.REF<вн> ] ⇒ X.Название

b. [ X.Rel.Y <ГГ1>, $t<ИГ, вн>, X.REF<им>] ⇒ Y.Название

*// решает задачу определения объектов метод прямого исключения рекурсии;*

c. [ X.Rel.Y<ГГ1>, gap ,X.REF<им>, $t<ИГ, вн>] ⇒ Y.Название

*// наилучшим образом обеспечивает такой метод доступ к информации о потребителе;*

d. [ $t<ИГ, вн>, X.Rel.Y<ГГ1>, gap, X.REF<им>] ⇒ Y.Название

*// количественную оценку уникальности совпадающих признаков позволяет получить вероятностный метод;*

e. [X.REF, gap , “,”, ‘который’<им>, gap, X.Rel.Y<ГГ1>, $t<ИГ, вн>] ⇒ Y.Название

*// прием цепных подстановок, который выполняет расчет величины влияния факторов в общем комплексе их воздействия на уровень совокупного финансового показателя;*

f. [$t<ИГ>, “,”, ‘который’<вн>, gap, X.Rel.Y<ГГ1>, gap, X.REF<им>] ⇒ Y.Название

*// проблеме снижения энергозатрат, которую позволяет решить цитратный метод.*

2) Действительное причастие

g. [X.REF, “,”, gap, X.Rel.Y<Прич, пе, дст>, $t<ИГ, вн>] ⇒ Y.Название

*// метод недоопределенных вычислений, эффективно решающий задачу удовлетворения ограничений в самой общей постановке;*

h. [X.Rel.Y<Прич, пе, дст>, $t<ИГ, вн>, X.REF] ⇒ Y.Название

*// устанавливающий взаимосвязи физических величин дидактический метод размерностей;*

3) Деепричастие

i. [X.Rel.Y<Деепр, пе, дст>, $t<ИГ, вн>, “,”, X.REF<им>] ⇒ Y.Название

*// выполняя поиск в подмножестве набора данных, алгоритм clarans.*

### 3.2.2. Пассивные конструкции

Трехчленная пассивная конструкция зеркально отражает активную конструкцию. В пассивной конструкции:

- субъект выражается творительным падежом,
- объект личной формы глагола выражается именительным падежом,
- предикат выражен формой глагола, соответствующей страдательному залогу.

1) Возвратный глагол

Описываются конструкции с возвратными глаголами, имеющими невозвратные корреляты, то есть образованными от невозвратных с помощью постфикса *-ся* (помечаются в словаре признаком *Рефл*). Предикат простого предложения или придаточного

определительного выражен личным глаголом или зависимым от вспомогательного слова инфинитивом в составе группы сказуемого:

ГГ2 = [<Глаг, Рефл>] [<Глаг, Вспом>,<Инф, Рефл>]

При этом онтологическое отношение представлено в словаре невозвратным коррелятом и/или возвратным дериватом. Ниже приведены варианты метапатернов для простого предложения (a-b, d-e) и сложного предложения с придаточным определительным (c).

a. [$t<ИГ, им>, X.Rel.Y<ГГ2,>, gap, X.REF<тв> ] ⇒ Y.Название

// *статическая задача решается известным методом нечетких когнитивных карт*;

b. [X.Rel.Y<ГГ2>, $t<ИГ, им>, X.REF<тв> ] ⇒ Y.Название

// *выполняется поиск логических закономерностей методами интеллектуального анализа данных*;

c. [$t<ИГ>,“,”, ‘который’<им>,gap, X.Rel.Y<ГГ2>, gap, X.REF<тв>] ⇒ Y.Название

// *динамическая задача, которая также решается методом нкк*;

d. [ $t <ИГ>, “,” , X.Rel.Y< Прич, Рефл, дст>, gap, X.REF<тв> ] ⇒ Y.Название

// *ряд практически важных задач по определению равновесных и кинетических параметров, решавшихся этим методом*;

e. [ X.Rel.Y < Прич, Рефл, дст>, gap , X.REF<тв>, $t<ИГ> ] ⇒ Y.Название

// *решающаяся таким методом задача удовлетворения ограничений в самой общей постановке*;

2) Страдательное причастие

f. [ $t<ИГ>, “,”, X.Rel.Y<Прич, пе, стр>, gap, X.REF<тв>] ⇒ Y.Название

// *проблемы классификации образов, решаемые применением нейронных сетей*;

g. [X.Rel.Y<Прич, пе, стр>, gap, X.REF, $t<ИГ> ] ⇒ Y.Название

// *решенная предложенным методом задача аппроксимации функций*;

f. [$t<ИГ>, X.Rel.Y<Кратк_Прич, пе, стр>, gap, X.REF<тв>] ⇒ Y.Название

// *задача персонализации решена методами коллаборативной фильтрации.*

### 3.2.3. Глагольные конструкции с управляемым предлогом

В ряде случаев в наименованиях онтологических отношений представлено сильное или слабое предложное управление глагола. В онтологии эксперимента большинство таких

наименований составляют возвратные глаголы с предлогом *решается_на, решается_в, сводится_к, используется_в, применяется_в, применяется_к.* Возможно и краткое страдательное причастие с предлогом *представлен_на.* Ниже представлены варианты метапатернов для простого предложения (a, d-g) и сложного предложения с придаточным определительным (b-c). В них предикат выражен личным глаголом, причастием или кратким причастием, либо зависимым от вспомогательного слова инфинитивом или кратким причастием в составе группы сказуемого:

ГГЗ = [<Глаг/Кратк_Прич>] [<Глаг, Вспом>,<Инф>][<Глаг, Вспом>,<Кратк_Прич>]

Основной особенностью соответствующих конструкций является подъем прямого объекта в позицию субъекта (пассивизация) при отсутствии реального субъекта-агенса и ввод в рассмотрение косвенного объекта. В этом случае семантический класс косвенного объекта (например, *Организация, Информационный_Ресурс*) определяет возможность обратной конструкции, в которой косвенный объект поднимается в позицию подлежащего, т.е. связь с предикатом реализуется беспредложно (a).

В данной версии языка шаблонов возможно лишь обобщенно представить глагольное управление. Следует отметить возможность отрыва предлога от управляющей лексемы. При этом используемый в шаблонах разрыв по-прежнему характеризуется строгим отсутствием в нем пунктуации, что позволяет уменьшить шум при извлечении.

a. [X.REF<им>, gap, X.Rel.Y<ГГЗ>, gap, <Предл>, $t<ИГ>] ⇒ Y.Название

// *аналитические материалы представлены на сайте минэкономразвития России, подход к данной проблеме описан в работе л. флорианя*;

Обратная конструкция:

[$t<ИГ,им>, X.Rel.Y<ГГЗ>, gap, X.REF<вн>] ⇒ Y.Название

// *сайт минэкономразвития представляет аналитические материалы*;

b. [X.REF, “,”, ‘который’<им>, gap, X.Rel.Y<ГГЗ>, gap, <Предл>, $t<ИГ>] ⇒ Y.Название

// *метод, который применяется только к крахмалосодержащим растениям*;

c. [$t<ИГ>, “,”, <Предл>, ‘который’, gap, X.Rel.Y<ГГЗ>, gap, X.REF<им>] ⇒ Y.Название

// *этап изучения объекта системного исследования, на котором возникает задача формализации объекта*;

d. [Предл, $t<ИГ>, X.Rel.Y<ГГ3>, gap, X.REF<им>] ⇒ Y.Название

// *на этапе анализа данных применяется не только метод размерностей*;

e. [ X.Rel.Y <ГГ3>, gap, X.REF<им>, gap, <Предл>, $t<ИГ>] ⇒ Y.Название

// *применяется описанный диахронический подход в ономасиологии*;

f. [X.Rel.Y<Прич>, gap , <Предл>, $t<ИГ, вн>, X.REF] ⇒ Y.Название

// *примененный в данном исследовании метод экспертных оценок;*

g. [X.REF, “,”, gap, X.Rel.Y<Прич>, gap , <Предл>, $t<ИГ> ⇒ Y.Название

// *Лоренцева ионизация, часто применяющаяся в физике ускорителей; методы квазиклассического приближения, применяемые в квантовой физике.*

### 3.3. Извлечение значений атрибутов

На основе типовых паттернов можно генерировать Т-ЛСП, предназначенные для извлечения значений атрибутов объектов класса X, которые в онтологии представлены свойством Datatype Property.

[X, X.A.type, $t<ИГ>] ⇒ X.A (3)

Во втором компоненте шаблона указано явно выраженное имя атрибута, а в третьем компоненте – извлекаемое свойство объекта ***$t,*** которое должно быть представлено именной группой. В случае ключевого атрибута *Название* это эквивалентно извлечению объекта. Результатом применения таких паттернов будет создание новых терминов и приписывание им семантического признака X.A.

В разделе 3.1 предложен подход к извлечению неизвестных базе знаний терминов на основе имен классов, представленных ИГ. Полезным дополнением этому является извлечение значений атрибута *X.Название* с учетом контекста, содержащего предикат класса *Именование* (*называть, именовать, название, под названием*) и явным образом говорящего о вводе в текст нового термина.

Соответствующие паттерны используют предварительно заданные шаблоны именной группы и глагольной лексемы, включающей личные формы глагола и причастия:

Г = [<Глаг>] [<Прич>] [<Кратк_Прич>]

***gap1*** обозначает возможный разрыв, не исключающий наличия запятой:

gap1 = [s/"."][s/"!"][s/"?"][s/":"][s/";"][s/"("][s/")"][s/"/””][s/"/n"]

Ниже представлены паттерны, выделенные при рассмотрении конструкций с терминами, представляющими наименования объектов различных классов (приведены примеры для классов X = *Метод*, *Научное_направление*, *Задача*, *Информационный_Ресурс*). Конструкции классифицируются в соответствии с грамматическим классом предиката наименования.

1) Глагольные конструкции

a. [X<им/вин>, gap1, Именование<Г> , gap, $t<ИГ, тв> ] ⇒ X.Название

*// Применяемый в работе метод **называется** методом магнетрона; Описывается метод, **называемый** методом согласования для обращения соответствующих интегральных уравнений; Используются различные математические методы, **именуемые** в данном контексте экономико-математическими*

b. [‘предлагать’<Г>, gap, X<им/вин>, gap1, Именование<Г>, gap, $t<ИГ, тв>] ⇒ X.Название

*// По аналогии с известным расчетным статическим методом предлагаемый метод **назовем** расчетным динамическим; В данной работе предлагается метод, **называемый** «элитным отбором» или «элитной стратегией»; Предлагаемый метод **назовем** методом двух групп; В работе предлагается метод, **называемый** Voxel Cone Tracing (VCT); В качестве альтернативного способа предлагается метод, **называемый** криотерапией;*

Снятие ограничения на падеж (тв) расширяет класс распознаваемых конструкций*:*

*// В работе предлагается метод, **называемый** иерархическая редукция (hierarchical reduction).*

Приведенные паттерны позволяют учесть и достаточно частотные в научных статьях конструкции с местоименным анафорическим элементом:

c. [‘предлагать’<Г>, gap, X<им/вин>, gap1, Именование<Глаг>, gap, <Мест, вн>, $t<ИГ, тв>] ⇒ X.Название

*// Предлагаемый метод, **назовем** его «методом перекоса»; Предлагаемый метод, **назовем** его «оценочное взвешенное пересечение».*

2) Конструкции с существительным

d. [X, gap1, <Глаг, Вспом>, Именование<Сущ> , gap, $t<ИГ> ] ⇒ X.Название

*// Имеется развитое направление исследований, **получившее название** математической экономики; Классическая задача исследования операций **получила название** обобщенной транспортной задачи; Один из подходов к многоэтапному оптимальному выбору в условиях*

*вероятностной неопределенности, который **носит название** дерева решений (decision tree); В рассказе и. одоевцевой, **носящем название** цитированного здесь стихотворения г. иванова «эпилог»; Явление, **получившее** в современных исследованиях **название** гипертекста; Проклятия у кадарцев **носят название** дерга; Вычислительное направление исследований в дальнейшем трансформировалось в новую методологию и технологию проведения научных исследований, которое **получило название** вычислительного эксперимента.*

3) Конструкции с лексическим оборотом *под названием*

e. [X, gap1, ‘под названием’, $t<ИГ>] ⇒ X.Название

*// kaufman и rousseeuw (1990) также предложили алгоритм, известный сейчас **под названием** clara (clustering large applications); после второй мировой войны стало развиваться научное направление **под названием** "исследование операций"; «принятие решений» — один из основополагающих терминов в научном направлении, известном **под названием** «исследование операций»; в основной статье в этом сборнике **под названием** "основы теории измерений", изложение шло на абстрактно-математическом уровне; Тянь чжан и др. предложили метод агломерационной иерархической кластеризации **под названием** birch; Ветви экономической теории, известной в официальных кругах **под названием** «статистика».*

Таким образом, из текста конкретного жанра с помощью характерных для данного жанра шаблонов извлекаются целевые понятия, зафиксированные в паттерне содержания.

## 5. Автоматическая генерация терминологических паттернов

Как было сказано выше, терминологические лексико-семантические паттерны автоматически строятся на основе онтологии паттернов онтологического проектирования, словаря общенаучной лексики и текущей версии онтологии НПО. На Рис.2 представлена схема взаимосвязей компонентов системы, на основе которых осуществляется генерация Т-ЛСП и пополнение словаря.

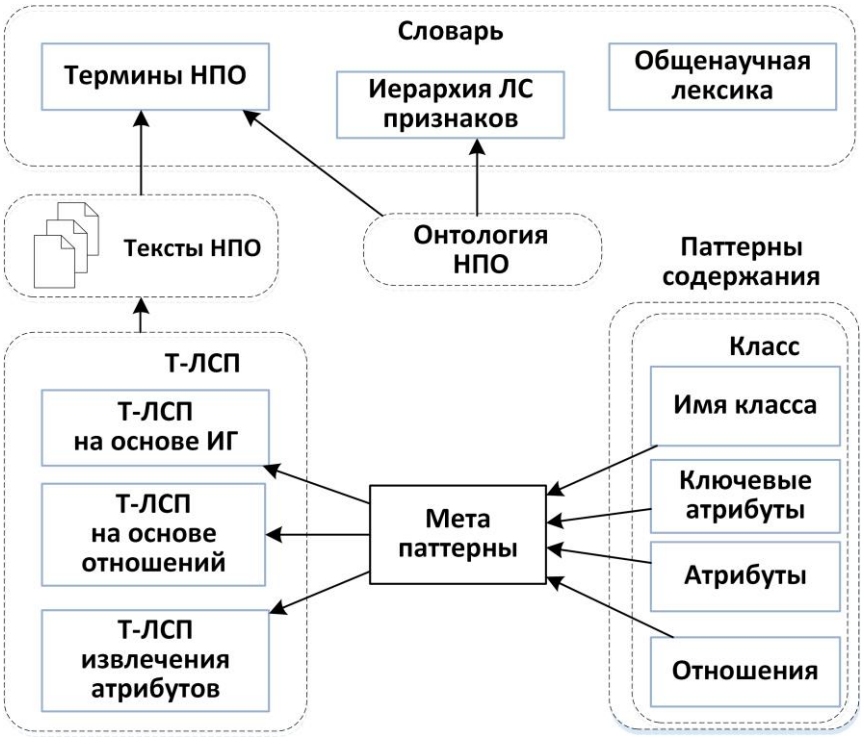


*Рис.2. Схема взаимосвязей компонентов системы знаний для генерации Т-ЛСП.*

Процесс генерации ЛСП начинается с создания и наполнения предметного словаря. Из онтологии и описания паттерна содержания извлекаются термины (лексемы и терминоподобные N-граммы) и формируются лексико-семантические классы по приведенной в п.1.2. методике. Все термины размечаются соответствующими семантическими признаками.

Отношение синонимии в явном виде отсутствует в словаре, однако, мы считаем квазисинонимами все термины, размеченные одним лексико-семантическим признаком. Поэтому для терминов с соотнесенными предметными и универсальными признаками можно выявить квазисинонимы по универсальному признаку (из словаря общенаучной лексики) и разметить их соответствующим предметным признаком.

На этапе генерации словаря используется корпус научных текстов, релевантных для рассматриваемой НПО, поскольку он является источником предметной лексики, употребляемой специалистами в данной НПО. Из корпуса можно извлечь не только типовые названия сущностей (классов сущностей), но и специфические индикаторы отношений или атрибутов.

На следующем этапе на основе анализа структуры паттерна содержания осуществляется означивание переменных в мета-паттернах и формируются Т-ЛСП. С помощью созданных Т-ЛСП анализируются тексты научного корпуса и словарь пополняется найденными новыми терминами.

При извлечении объектов необходима информация о ключевых атрибутах классов онтологии. В приведенных выше примерах в качестве такого атрибута выступает *Название*, однако у класса могут быть и другие ключевые атрибуты (например, *Фамилия* и *Имя* у *Персоны*) которые потребуют создания других типовых паттернов.

Таким образом, на основе онтологических компонентов знаний можно выделить необходимые для извлечения информации из текста и пополнения онтологии знания о языке предметной области и возможные способы языкового описания онтологических сущностей в текстах. Формализация этих знаний в виде системы ЛСП позволяет применить существующие технологии автоматической обработки текстов для решения поставленной задачи.

## 5. Анализ экспериментов

На основе предложенной методики проведен эксперимент по созданию Т-ЛСП и извлечению новых терминов для онтологии НПО «Поддержка принятия решений в слабоформализованных областях». Экспериментальная проверка была проведена для паттерна содержания *Метод*. На основе текущего состояния онтологии данной НПО и описания паттерна содержания *Метод* были автоматически созданы словарь предметной области, включающий 214 терминов, размеченных с помощью 21 лексико-семантического признака, и 34 Т-ЛСП для извлечения новых терминов (названий экземпляров классов и предикатных слов). Для генерации иерархии лексико-семантических признаков использовались метки (rdfs:label) атрибутов и отношений класса *Метод*, а также значения атрибутов для экземпляров этого и связанных с ним классов.

Для оценки качества полученных ЛСП использовался корпус, состоящий из научных публикаций на русском языке, общим объемом 53,9 тыс. токенов. С помощью Т-ЛСП, предназначенных для извлечения наименований индивидов класса *Метод*, найдено 111 вхождений и выделено 73 уникальных термина, которым был приписан лексико-семантический класс *Метод.Название*. Из них 70 терминов оказались новыми (отсутствовавшими в словаре, построенном по онтологии): *метод анкетного опроса, метод медиан рангов, метод нейронных сетей, метод анализа иерархий Саати, метод интервью, метод самооценки* и т.д. Аналогичные Т-ЛСП применялись для извлечения имен индивидов других классов, связанных с классом *Метод*. Так, для класса *Задача* было найдено 42 вхождения шаблона и выявлено 33 термина, из которых 30 являлись новыми: *задача когнитивного моделирования*, *задача порядковой классификации*, *задача целочисленного программирования* и т.д. Среди ошибочных результатов можно отметить термины с

указанием источника, строящихся по тому же синтаксическому правилу, например, *задача дипломной работы*.

Для извлечения новых предикатных терминов использовались Т-ЛСП, построенные по мета-паттернам, моделирующим ситуацию онтологической связи двух объектов. Так, для Т-ЛСП, служащего для извлечения предикатных терминов, относящихся к классу *Метод.решает.Задача*, в корпусе было найдено 22 вхождения и выявлены такие термины как: *поставили, позволяют построить, позволил выделить, ищут решения, позволяет решать* и т.п. Всего было найдено 38 новых предикатных терминов.

Общая точность работы Т-ЛСП была оценена тремя экспертами и составила в среднем 73,97%, при этом согласие между экспертами оценено в 81,74%.

Анализ материалов и результатов эксперимента выявил недочеты в описаниях контекстов извлечения, приводящих к ложно-негативным или ложно-позитивным результатам.

1) Использование наименований классов в качестве индикаторов имен соответствующих индивидов имеет ряд ограничений а) формирование имени индивида на основе имени класса покрывает лишь некоторое подмножество упоминаемых в текстах индивидов; б) далеко не всегда для конкретного класса в качестве индикатора, т.е. вершины именной группы, представляющей наименование индивида, используется модель с полным именем класса – чаще используется голая вершина или альтернативная модель, в которой имя класса подчиняет зависимые, не учтенные в модели именования индивидов 1 типа:

*деятельность* + *по* + Сущ vs. **деятельность* + Сущ,рд

2) Морфолого-синтаксическая информация используется неполно в силу специфики языка шаблонов: а) невозможно задать общие условия согласования морфологических характеристик существительного и зависимого прилагательного (можно лишь задать характеристики препозитивного прилагательного, если определены соответствующие характеристики существительного); б) невозможно задать общие условия соответствия морфологических характеристик подлежащего и сказуемого; в) невозможно учесть падежное управление индивидуальных глаголов с помощью общего ограничения (последние два ограничения можно учесть лишь частично для определенных конструкций, таких как активная конструкция для переходных глаголов и соответствующая пассивная конструкция на базе страдательных форм и возвратных глаголов).

3) Методика поверхностного (неполного) синтаксического анализа, реализуемого с помощью языка шаблонов, не всегда позволяет решить важную проблему границ покрытого текстового фрагмента, прежде всего, границ именной группы (что характерно для всей области извлечения информации).

4) Разрывный контекст, покрываемый ЛСП, представляет существенную проблему: так, необходимо исключить из контекстов не только лексическое отрицание, но и другие лексические и грамматические способы выражения ирреальности и проблематичной достоверности (сомнения): *не может, не позволяет, едва ли, сомнительно* и т.п.

5) Особенности исходной онтологии определяют формулировки генерируемых ЛСП: а) имеющиеся в онтологии орфографические или содержательные ошибки, такие как неверное задание значения атрибута (описание вместо названия) или отнесение индивида к классу (название организации в классе объектов исследования), определяют генерацию некорректных ЛСП; б) генерация ЛСП затруднена отсутствием стандартного языка представления онтологических сущностей и отношений. В качестве примеров можно привести именные группы сложной структуры (в том числе сочинительные конструкции) в названиях классов, использование нумерации в названии объектов.

## Заключение

В статье представлен основанный на правилах лингвистический подход к извлечению из текстов информации для поддержки процесса автоматического пополнения онтологии на основе исходной онтологии, начального (универсального) словаря и корпуса текстовых документов. Принятый подход базируется на следующих методических принципах:

- автоматическое пополнение предметного словаря на основе онтологии и корпуса текстов и разметка его с помощью системы семантических признаков, также основанных на онтологии,
- определение небольшого набора исходных структурных мета-паттернов, устанавливающих концептуальные контексты извлечения онтологической информации,
- автоматическое порождение по набору структурных мета-паттернов множества лексико-семантических паттернов, определяющих лексические, семантические и синтаксические свойства контекстов извлечения.

В работе приведены основные паттерны для извлечения объектов (экземпляров классов онтологии) и их свойств с учетом вариативности и множественности языковых выражений, представляющих элементы онтологии в тексте на русском языке, отличающемся сильной флективностью и свободным порядком слов.

Данный подход апробируется при разработке и пополнении онтологий различных научных предметных областей («Поддержка принятия решений», «Поддержка решения вычислительно сложных задач на суперкомпьютерах»).

# Список литературы


1. Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны в задачах автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог’2007. М.: Издательский центр РГГУ. 2007. С. 70-75.
2. Жигалов В.А., Жигалов Д.В., Жуков А.А., Кононенко И.С., Соколова Е.Г., Толдова С.Ю. Система Alex как средство для многоцелевой автоматизированной обработки текстов // Труды международного семинара Диалог’2002 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука. 2002. Т.2. С. 192-208.
3. Загорулько Ю.А., Сидорова Е.А., Загорулько Г.Б., Ахмадеева И.Р., Серый А.С. Автоматизация разработки онтологий научных предметных областей на основе паттернов онтологического проектирования // Онтология проектирования. 2021. Т.11, №4(42). С. 500-520.
4. Ковалев А.И., Сидорова Е.А. Инструмент разработки предметных словарей на основе лексических шаблонов DigLex // Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ–2015), 6 - 8 октября 2015 г., Новосибирск. Новосибирск: Институт математики им. С.Л. Соболева СО РАН. 2015. Т. 1. С. 123–130.
5. Рабчевский Е.А. Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска // Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL’2009. Петрозаводск. 2009. С. 69–77.
6. Alani H., Kim S., Millard D.E., Weal M.J., Hall W., Lewis P.H., Shadbolt N.R. Automatic Ontology-Based Knowledge Extraction from Web Documents // IEEE Intelligent Systems 18(1), 2003. P.14–21.
7. Akhmadeeva I. R., Zagorulko Y. A., Mouromtsev D. I. Ontology-Based Information Extraction for Populating the Intelligent Scientific Internet Resources // Knowledge Engineering and Semantic Web: 7th International Conference, KESW 2016, Proceedings. / Ngomo A. C. N., Křemen P. (Eds.). Communications in Computer and Information Science. Springer International Publishing, 2016. Vol. 649. P. 119-128.
8. Blomqvist E., Hammar K., Presutti V. Engineering Ontologies with Patterns: The eXtreme Design Methodology // In: Hitzler P., Gangemi A., Janowicz K., Krisnadhi A., Presutti V. (eds.): Ontology Engineering with Ontology Design Patterns. Studies on the Semantic Web. Amsterdam, IOS Press, 2016. Vol. 25. P. 23–50.
9. Ganino G., Lembo D., Mecella M., Scafoglieri F. Ontology population for open-source intelligence: a GATE-based solution // Software: Practice and Experience. 2018. 48(12). P. 2302-2330.
10. Hearst M.A. Automatic Acquisition of Hyponyms from Large Text Corpora // Proceedings of the 14th International Conference on Computational Linguistics. 1992. P. 539–545.

11. Ijntema, W., Sangers, J., Hogenboom, F., Frasincar, F. A lexico-semantic pattern language for learning ontology instances from text // Journal of Web Semantics. 2012. Vol. 15. P. 37–50.
12. Maynard D, Funk A, Peters W. Using Lexico-Syntactic Ontology Design Patterns for Ontology Creation and Population // In: Proc. Workshop on Ontology Patterns (WOP 2009), collocated with the 8th Int. Semantic Web Conf. (ISWC-2009). CEUR Workshop Proceedings. 2009. Vol. 516. P. 39–52.
13. Petasis G, Karkaletsis V, Paliouras G, Krithara A, Zavitsanos E. Ontology Population and Enrichment: State of the Art. In: Paliouras, G., Spyropoulos, C.D., Tsatsaronis, G. (eds): Knowledge-Driven Multimedia Information Extraction and Ontology Evolution). Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2011. Vol. 6050. P. 134–166.
14. Ruiz-Martínez M.J., Valencia-García R., Martínez-Béjar R., Hoffmann A. 2012 BioOntoVerb: A top level ontology based framework to populate biomedical ontologies from texts // Knowledge-Based Systems. 2012. Vol. 36. P. 68-80.
15. Saeeda L., Med M., Ledvinka M., Blaško M., Křemen P. Entity Linking and Lexico-Semantic Patterns for Ontology Learning // The Semantic Web (ESWC 2020). Lecture Notes in Computer Science. Springer, Cham. 2020. Vol. 12123. P. 138–153.
16. Zagorulko Yu. A., Zagorulko G.B., Borovikova O.I. Pattern-Based Methodology for Building the Ontologies of Scientific Subject Domains // In: H. Fujita and E. Herrera-Viedma (Eds.): New Trends in Intelligent Software Methodologies, Tools and Techniques. Proceedings of the 17th International Conference SoMeT_18. Series: Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press, 2018. Vol. 303. P. 529-542.