УДК 004.8

*Палагин А.В.* [а], *Петренко Н.Г.* [а], *Зеленцов Д.Г.* [б]

# К ВОПРОСУ КОМПЬЮТЕРНОЙ ОБРАБОТКИ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

[а] **Институт кибернетики имени В.М. Глушкова НАН Украины, г. Киев, Украина**
[б] **ГВУЗ «Украинский государственный химико-технологический университет», г. Днепр, Украина**

В статье рассмотрен общий подход к проблеме анализа естественно-языковой информации, включающий реализацию ряда информационных технологий, тем или иным образом связанных с языковым моделированием. Кроме разработки указанных информационных технологий, необходимо разработать формальную теорию компьютерной обработки знаний, извлеченных из естественно-языковых текстов. Проанализированы особенности построения лингвистических моделей и критерии понимания естественно-языковых текстов. При этом возникает ряд проблем. Первая проблема сводится к проблеме анализа текстовой информации, представленной на естественном языке (морфологический, синтаксический, семантический и логический анализ) с целью извлечения знаний. Вторая проблема связана с проектированием системы поиска, обработки и извлечения знаний, разработки и построения ее архитектуры, а также инструментария для пользователя. И третьей проблемой является разработка процедур интеграции знаний из нескольких предметных областей для обеспечения эффективности проведения исследований междисциплинарного и трансдисциплинарного характера. Также необходимо уделить особое внимание вопросам использования уже наработанных теоретических положений и практических решений. Предложена формальная постановка задачи анализа естественно-языковых текстов, в которой выделены основные подзадачи, связанные с вычислением отношений типизации лексики естественного языка на лексико-смысловом континууме и интерпретацией некоторого текста на заданной предметной модели. В контексте разработанной архитектуры языково-онтологической информационной системы предложена формальная модель обработки естественно-языковых текстов, для которой показано однозначное соответствие процессов обработки естественно-языковой информации и средств (архитектурных блоков) их реализации.

**Ключевые слова:** ELRE естественно-языковый текст, лингвистическая модель, языково-онтологическая информационная система, анализ и понимание естественно-языковых текстов.

**DOI:** 10.32434/2521-6406-2020-1-7-37-45

***Постановка проблемы***

Степень развития и внедрения компьютерных технологий в значительной мере определяется не столько возможностями традиционной вычислительной техники, сколько особенностями предметных областей, и успех эффективной компьютеризации последних существенным образом зависит от глубины исследования и понимания моделируемого явления. Исходя из этого этап формирования информационных технологий (ИТ) исключительно с позиций потенциальных возможностей вычислительной техники постепенно смещается в сторону дифференциации отдельных направлений ИТ, где на первый план выступают особенности конкретных предметных областей (ПдО). К такому направлению следует отнести ИТ, которые тем или иным образом связаны с языковым моделированием поведения человека, или системы, ориентированные на обработку естественно-языко-

© Палагин А.В., Петренко Н.Г., Зеленцов Д.Г., 2020

вой информации. Причем под обработкой естественно-языковых текстов (ЕЯТ) понимается реализация ряда ИТ, конечной целью которых является компьютерная обработка знаний как высшей формы умственной деятельности человека.

Разработка формальной теории компьютерной обработки знаний составляет одну из насущных проблем в общей теории искусственного интеллекта. Сложность указанной проблемы определяется, в частности, необходимостью привлечения целого ряда научных теорий (математической логики, компьютерной и психологической лингвистики, нейрофизиологии, нейрокибернетики и др.), которые в совокупности, будучи примененными к решению проблемы формального представления и обработки знаний, составили бы концептуально единую (междисциплинарную) формальную теорию. Составляющие указанной теории должны учитывать сущность этапов языковой и предметной обработки информационных объектов (для первого – ЕЯТ, а для второго – извлеченные из ЕЯТ знания) и формальной связи между ними. В конечном итоге необходимо разработать комплекс информационных технологий компьютерной обработки ЕЯТ, представления знаний и их компьютерной обработки. При этом возникает несколько проблем [1–3].

Первой проблемой (и одной из основных) является проблема анализа текстовой информации, представленной на естественном языке (морфологический, синтаксический, семантический и логический анализ) с целью извлечения знаний.

Второй проблемой является проблема проектирования системы поиска, обработки и извлечения знаний, разработки и построения ее архитектуры, а также инструментария для пользователя.

Третьей проблемой является проблема интеграции знаний из нескольких предметных областей для обеспечения эффективности проведения исследований междисциплинарного характера, использования уже наработанных теоретических положений и практических решений.

***Цель статьи***

Целью статьи является разработка формальных моделей информационного процесса компьютерного анализа и понимания естественно-языковых текстов и архитектуры системы его реализации.

***Анализ последних исследований и публикаций***

При решении указанных проблем (и разработки формальной теории) важной задачей является построение естественно-языковых лингвистических моделей и создание на их основе эффективных лингвистических процессоров, которые в совокупности с языково-онтологической картиной мира (ЯОКМ) представляют языково-онтологическую информационную систему.

Лингвистические модели – это, в сущности, фундаментальная научно-прикладная область исследований, помогающая строить системы обработки ЕЯТ. Под последней понимается процесс взаимодействия «Система-ЕЯТ-Пользователь», который включает в себя разные способы взаимодействия с ЕЯТ, такие как анализ, генерация, интерпретация, трансформация, синтез и др. Такое определение лингвистических моделей, основанное на их функциональном аспекте, является полезным с методологической точки зрения, обеспечивая классификацию моделей по их прагматическим признакам, т.е. по цели разработки и сфере применения. Выделим следующие классы лингвистических моделей:

1) диалоговые “запрос-ответ” или интерактивные модели;

2) концептуально-формальные модели;

3) концептуально-функциональные модели;

4) когнитивные (семантико-контекстные) модели.

Очевидно, наиболее сложными являются последние модели. Именно они обеспечивают глубинное проникновение в текущий контекст и его трансформацию с сохранением смысла как внутри одной модели, так и между разными моделями. При этом объектом исследования являются естественно-языковые тексты, представленные в электронном виде и взятые из сети Интернет, монографий, учебников, научно-технических документов и т.п., которые в совокупности составляют исходный лингвистический корпус текстов.

В существующих интеллектуальных системах (ИС) выделяют пять основных уровней понимания ЕЯТ [4].

Первый уровень. Характеризуется схемой, показывающей, что любые ответы на вопросы система формирует только на основе прямого содержания, вытекающего из текста. В лингвистическом процессоре выполняется морфологический, синтаксический и семантический анализ текста и вопросов, относящихся к нему. На выходе лингвистического процессора получаем внутреннее представление текста и вопросов, с которыми может работать блок вывода. Он фор-

мирует ответы, используя специальные процедуры. Другими словами, уже понимание на первом уровне требует от ИС определенных способов представления данных и вывода на этих данных.

Второй уровень. На втором уровне добавляются способы логического вывода, основанные на информации, содержащейся в тексте. Это различные логики текста (временная, пространственная, каузальная и др.), порождающие информацию, явно отсутствующую в тексте. Архитектура ИС, с помощью которой может быть реализован второй уровень понимания, должна иметь дополнительную базу знаний, в которой хранятся закономерности, относящиеся к временной структуре событий, возможной их пространственной организации, каузальной зависимости и т.д., а логический блок – все необходимые средства для работы с неклассическими логиками.

Третий уровень. К средствам второго уровня добавляются правила пополнения текста знаниями системы о среде. Эти знания в ИС, как правило, носят логический характер и фиксируются в виде сценариев или процедур другого типа. Архитектура ИС, в которой реализуется понимание третьего уровня, внешне не отличается от архитектуры ИС второго уровня. Но в логическом блоке должны быть учтены средства не только для чисто дедуктивного вывода, а и для вывода по сценариям.

Три перечисленные уровня понимания полностью или частично реализованы практически во всех действующих ИС.

Четвертый уровень. На этом уровне происходит изменение содержимого базы знаний. Она дополняется фактами, известными системе и содержащимися в тех текстах, которые введены в систему. Разные ИС отличаются одна от другой характером правил порождения фактов из знаний, опираясь на методы дедуктивного вывода и распознавания образов. Правила могут быть основаны на принципах вероятностей, размытых выводов и т.д. Но во всех случаях база знаний оказывается априорно неполной. В ИС возникают сложности с поиском ответов на запросы. В частности, в базах знаний становится необходимым немонотонный вывод.

Пятый уровень. На этом уровне происходит порождение метафорического знания. Правила порождения знаний метафорического уровня, используемых для этих целей, представляют собой специальные процедуры, опирающиеся на выводы по аналогии и ассоциации. Известные в настоящее время схемы вывода по аналогии используют, как правило, диаграмму Лейбница, которая отображает только частный случай суждений по аналогии. Еще меньше разработаны схемы ассоциативных суждений.

Существуют и другие интерпретации феномена понимания. Возможно, например, оценивать уровень понимания по способности системы к пояснению полученного результата. Здесь возможен не только уровень пояснения, когда система поясняет, что она сделала, например, на основе введенного в нее текста, а и уровень обоснования (аргументации), когда система обосновывает свой результат, показывая, что он не противоречит той системе знаний и данных, которыми она владеет. В отличие от пояснения обоснование всегда связано с суммой фактов и знаний, которые определяются текущим моментом существования системы. И введенный для понимания текст в одних состояниях может быть воспринят системой как истинный, а в других – как ложный. Существующие ИС типа экспертных систем, как правило, способны давать пояснения и лишь частично обоснования.

Обобщенный критерий понимания ЕЯТ научно-технического профиля состоит в способности решать прикладные задачи на основе содержащихся в них знаний.

Особенности анализа ЕЯТ определяются направленностью на формирование структуры понятий, то есть, на автоматическое извлечение знаний из текстов и их прагматическую интерпретацию в терминах прикладной задачи. При этом текст рассматривается как объект разных уровней анализа: как знаковая система, как грамматическая система и как система знаний про ПдО. Каждый уровень имеет свои особенности, свои способы выражения и, следовательно, допускает наличие специфических методов обработки.

***Изложение основного материала***

*Общие принципы анализа и понимания ЕЯТ.* Исследование процессов интерпретации и понимания языковых высказываний, имеет как теоретический, так и прикладной интерес. Работы в области автоматического анализа текста и автоматического решения задач, сформулированных на естественном языке (или языке, близким к нему), показали актуальность построения интерпретационной теории языка. Понятие “интерпретация” с самого начала лежало в основе общелингвистических теорий, а также в основе логических исследований. Интерпретационный

подход представлен как в разнообразных областях чисто лингвистического анализа (в теориях формальных грамматик, в "теории языковых актов" и т.п.), так и в исследованиях по искусственному интеллекту. Согласно интерпретации в основе владения языком и его использованием лежит один и тот же интерпретирующий механизм, обслуживающий разные сферы языковой деятельности и использующий разные виды знаний. Среди этих сфер – речь, понимание, редактирование, комментирование, перефразировка, соображение, коммуникация, аргументация, обучение, перевод и др. Сама же интерпретация, через которую и определяются указанные сферы, представляет собой получение на основе одного исходного объекта (объекта, который интерпретируется) другого, отличного объекта, который допускается интерпретатором как равнозначный исходному на фоне конкретной ситуации, набора знаний [5]. Сами же знания не входят в структуру языка непосредственно, а "привлекаются" к интерпретации, и только опосредованно определяют результат интерпретации языковых высказываний.

Одной из основных процедур обработки ЕЯТ является процедура распознавания, в частности когнитивного распознавания. Под когнитивным распознаванием ЕЯТ понимается процесс формализации извлечения и представления знаний предметной области, содержащихся в ЕЯТ. Входом процедуры распознавания является ЕЯТ, а выходом (результатом) – формально-логическое представление. Оно является формализованным представлением знаний о ПдО, отраженным в определенном ЕЯТ.

Процесс распознавания и извлечения знаний из ЕЯТ базируется на компьютерном моделировании интеллектуальных функций человека, а именно – на моделировании процесса понимания человеком ЕЯТ. При этом термин понимание определяется с помощью таких критериев: умение пересказать "своими" словами, т.е. другими (лексическими, синтаксическими) средствами передать содержание входного текста, умение ответить на вопрос относительно определенного текста и др. Процедура распознавания базируется на средствах формализации (т.е. разработки онтологических моделей) знаний об определенном языке и знаний об определенной ПдО.

Отметим, что когнитивное распознавание ЕЯТ имеет свои особенности, а именно:

– внеязыковые ситуации, описанные в текстах, определяются лишь знаниями из определенной ПдО;

– средства вербализации этих знаний ориентированы на определенный уровень профессиональной подготовки;

– механизмы взаимодействия знаний в тексте с когнитивной картиной мира основываются на модели представления человеком знаний об определенном языке (языковая картина мира [6]) и знаниях о фрагментах реальной действительности (базы знаний ПдО).

*Формальная постановка задачи анализа ЕЯТ[7].* Пусть $T=t_1, t_2, ..., t_n$ естественно-языковый текст в алфавите X, т.е. $T\in L(X)$, где $L(X)$ – язык над алфавитом X, а $t_i\in T$ – предложения, $i=\overline{1,n}$, n – мощность множества T.

Каждое предложение $t_i\in T$, в свою очередь, имеет структуру $t_{i_1}, t_{i_2}, ..., t_{i_m}$, где $t_{i_j}$ содержательно означают грамматические единицы, из которых построено предложение $t_i$. Если $t_{i_j}\in t_i$, то $C_L\left(t_{i_j}\right)=t_{i_1}, t_{i_2}, ..., t_{i_{(j-1)}}$ и $C_R\left(t_{i_j}\right)=t_{i_{(j+1)}}, t_{i_{(j+2)}}, ..., t_{i_m}$ будем называть левым и правым контекстом $t_{i_j}$ соответственно.

С текстом T связаны такие объекты:

– S – словарь языка L(X), где содержатся слова $t_{i_j}$ со своими определителями (в частности, лингво-семантическими характеристиками единиц словаря);

– $\gamma\subseteq T\times S$ – отношение, определяющее возможные значения и типы слова в словаре S;

– A=(D,П) – предметная модель, на которой интерпретируется текст T;

– $\phi\subseteq T\times A$ – отношение интерпретации текста T на области D.

Опуская формальные преобразования, отметим, что из этой постановки задачи анализа ЕЯТ вытекают следующие основные подзадачи:

– конкретизировать предметную модель A; задача связана с тем, что предметная модель является по существу базой знаний (конкретизация состоит в том, чтобы определиться с формальным логическим языком, правилами вывода, аксиоматикой и пр.);

– показать вычислимость отношений $\gamma$ и $\phi$ на предметной модели A;

– построить алгоритмы вычисления отношений $\gamma$ и $\phi$;

– при вычислении отношений $\gamma$ и $\phi$ контролировать соответствия типов аргументов и предикатов;

– определить взаимодействие алгоритмов вычисления $\gamma$ и $\phi$ с системами лингвистическо-

го анализа текста.

Подзадачами второй очереди являются подзадачи, связанные с:

– определением структуры данных и информации для словарей;

– определением режима взаимодействия с пользователем (автоматический, полуавтоматический или диалоговый).

*Лексико-грамматический анализ.* Пусть L – язык отношений, которые представляют знания, V – множество грамматических характеристик, включая грамматические разряды естественного языка (ЕЯ), а D – область интерпретации.

Лексико-грамматический анализ приводит к конкретизации интерпретации $\varphi$: $V \to D$ и отношений $R_i \in L$. Интерпретация $\varphi$ в данном случае представляет собой суперпозицию двух функций $\varphi_1$ и $\varphi_2$, т.е. $\varphi(V)=\varphi_2(\varphi_1(V))=\varphi_1*\varphi_2(V)$, где * обозначает суперпозицию функций. Функции $\varphi_1$ и $\varphi_2$ реализуют процесс синтаксического и семантического анализа предложений текста T, а отношения $R_1$ и $R_2$ – это синтаксические (правила языка, на котором написан текст T) и семантические ограничения.

Одним из дальнейших возможных уточнений является уточнение отображения $\varphi_1$. Это отображение, в свою очередь, можно рассматривать как суперпозицию двух отображений, реализующих морфологический и синтаксический анализ предложений ЕЯТ и вместе с отображением $\varphi_2$ образующих целостную систему классического типа, схема которой показана на рис. 1 [8]. Возможная структура словарей, которые используются в приведенной схеме, и некоторое ее обоснование описаны в работах [9].

*Информационная модель этапов лингвистического анализа.* Архитектура современных знание-ориентированных информационных систем с естественно-языковым представлением и обработкой знаний включает онтологическую составляющую эксплицитно, которую в общем виде можно интерпретировать как концептуальную базу знаний. Такая база знаний представляется в виде ориентированного графа, вершинами которого являются концепты, а дугами – множество отношений, связывающих между собой концепты. Другой важной особенностью указанной архитектуры является разделение и отдельная обработка семантики первой и второй ступени [10], что в общем случае означает разделение внутриязыкового и внеязыкового (экстралингвистического) процессинга и переход к формально-логическому представлению исходного текста.

Указанные особенности архитектуры современных знание-ориентированных информационных систем трансформируют традиционную модель обработки ЕЯТ в формальную модель следующего вида [7]:

$$F=\langle T,W,SS^1,O,S^2,I\rangle,$$

где T – множество обрабатываемых ЕЯТ; W – множество словоформ, входящих в T; $SS^1$ – множество синтактико-семантических структур первой ступени, описывающих T; O – множество онтологических структур, отображающих множества W и $SS^1$ в $S^2$; $S^2$ – множество семантических структур второй ступени, описывающих множество сценариев T; I – множество информационно-кодовых представлений $S^2$.

Опишем объекты формальной модели. Множество T представляет совокупность ЕЯТ, характеризующихся стилями делового и научно-технического характера.

Цепочка $W \to SS^1$ в классическом понимании представляет грамматический анализ ЕЯТ. В отличие от традиционных линейного и сильно кодированного методов анализа, здесь использован смешанный метод анализа. Суть его состоит в том, что в лексикографической базе

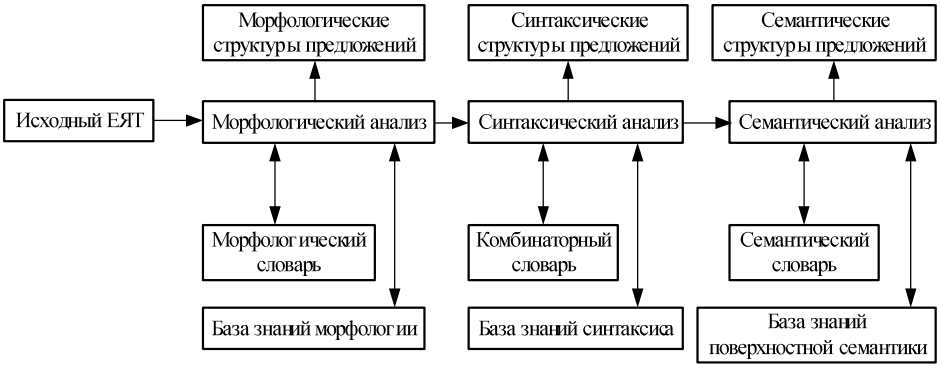


Рис. 1. Схема лингвистического анализа ЕЯТ классического типа

данных полное множество W представлено в таблицах двух типов: таблицах лексем с соответствующими морфологическими, синтаксическими и семантическими характеристиками и таблицах флексий для всех полнозначных изменяющихся частей речи. При этом алгоритмы формирования парадигмы лексем просты: в таблицах лексем указаны основы лексем и соответствующие коды для выбора записей из таблиц флексий. Нефлексийные изменения учитываются соответствующими алгоритмами.

Множество O онтологических структур в идеале представляет ЯОКМ.

Цепочки преобразования информации $T \rightarrow W \rightarrow SS^1$ и $O \rightarrow S^2 \rightarrow I$, по сути, представляют соответственно базовые процедуры анализа и понимания ЕЯТ, средствами интерпретации которых являются грамматический и семантический процессоры.

Практическая ценность получаемых результатов при обработке ЕЯТ, в основном, зависит от полноты интерпретационных моделей семантических структур ЕЯТ и их формального представления. Под полнотой понимается включение в модель как составной семантики первой ступени (или объектной составляющей), так и составной семантики второй ступени (или акторной составляющей). Такое распределение семантики хорошо согласовывается как с онтологической иерархией концептуальных категорий, так и со сложностью выполнения вычислительных процедур при компьютерной обработке ЕЯТ.

С точки зрения лингвистики, семантическая составляющая первой ступени описывается на уровне грамматики отдельных частей речи, в то время как составляющая второй ступени уже описывается синтаксическими конструкциями таких единиц синтаксиса, как предложение, абзац, параграф, раздел и текст. С точки зрения математической логики, если первую ступень можно описать (довольно условно) исчислением высказываний, то вторая ступень должна описываться исчислением предикатов с квантифицированными переменными.

Наибольшей полноты (и соответственно наибольшей степени сложности) приобретают модели, которые описывают ЕЯТ в целом. Такие модели описывают, в частности, некоторый сценарий, отображающий содержание ЕЯТ. В свою очередь, как ЕЯТ делится на синтаксические единицы, так и общий сценарий раскладывается на отдельные сценарии, ситуации и элементарные ситуации.

Описанное важное различие между объектной и акторной составляющими семантики, а также морфолого-синтаксическим анализом, в частности, в сложности их интерпретационных моделей, обусловила выделение для моделирования и интерпретации семантики отдельного функционального модуля – семантического процессора. Морфологический и синтаксический анализ при этом выполняется грамматическим процессором, а точнее отдельными его блоками (морфологического и синтаксического анализа). Он содержит также лингвистическую СУБД реляционного типа и синтаксическую базу знаний.

*Архитектурная организация ЯОИС [11,12].* Архитектура ЯОИС, разработанная в соответствии с моделью (1) для предметной области обработки ЕЯТ, представлена на рис. 2. На нем приняты следующие обозначения: ЛБД – лексикографическая база данных; ЯОКМ – языково-онтологическая картина мира.

ЛБД представляет набор таблиц, соответствующих грамматическому словарю для каждой части речи ЕЯ, таблиц падежных окончаний для формирования словоформ лексемы, уникальных идентификаторов лексем ЕЯ и их синтаксических и семантических характеристик. Все лексические единицы в таблицах соответствующим образом проиндексированы и имеют одинаковое интерпретационное значение, как для грамматического, так и семантического процессора. Заметим, что функции ЛБД значительно расширены по сравнению с традиционными грамматическими словарями, и они эффективно реализуются аппаратными средствами.

ЯОКМ представляет лингвистическую онтологию, одну из центральных компонент ЯОИС и совместно с ЛБД представляет базу знаний лексики ЕЯ. ЯОКМ – это формализованная онтология, в которой аксиомы и определения входят в состав баз знаний синтаксиса и семантики модуля грамматического анализатора (на рисунке не показаны), а ограничения являются составной частью синтаксических и семантических характеристик ЛБД.

Грамматический анализатор – это компонента, реализующая процедуры графемного, морфологического и синтаксического анализа. Она взаимодействует с лексикографической базой данных, а результаты анализа (сформированные в виде итоговой морфологической таблицы текста и синтаксических деревьев предложений, входящих в текст) передаются на вход семантического анализатора.

Семантический анализатор – это компо-

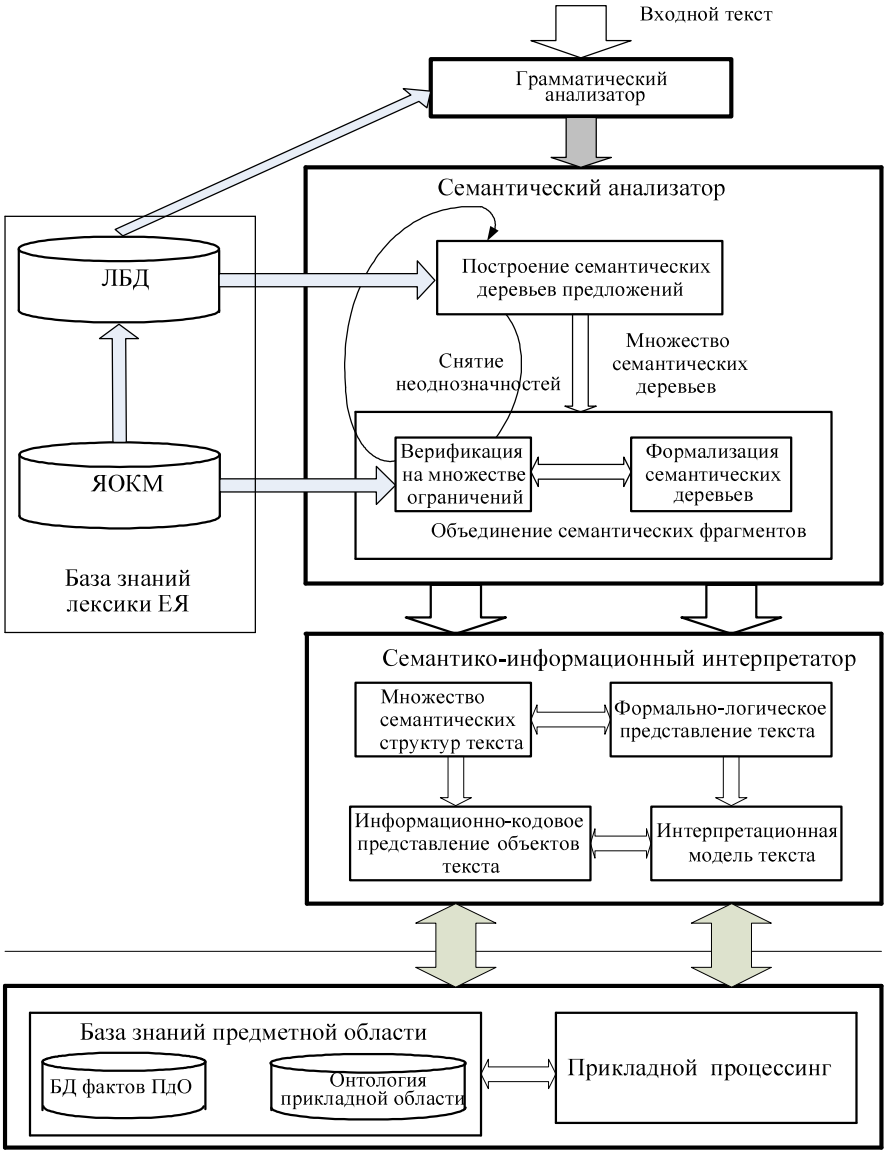


Рис. 2. Архитектура ЯОИС

нента, реализующая процедуры семантического анализа предложений текста, решения задачи грамматической и лексической неоднозначности и построения формально-логического представления предложений текста. При этом первые две процедуры могут выполняться итерационно. Семантический анализатор взаимодействует с ЛБД и ЯОКМ, а результаты анализа (сформированные деревья и формально-логические представления семантически связанных фрагментов текста) передаются на вход семантико-информационного интерпретатора.

Семантико-информационный интерпретатор реализует процедуры построения информационно-кодового представления семантики текста и его интерпретационной модели (знание-ориентированной компоненты). Причем, если построение первой составляющей обеспечивает прикладной процессинг непосредственно входного текста (реферирование, классификацию, построение простой онтологии документа и др.), то в совокупности со второй составляющей обеспечивается прикладной процессинг для различных процедур обработки не полностью формализованных знаний (извлечение, интеграция, накопление знаний и др.).

Модули прикладного процессинга и базы знаний предметной области предназначены для решения конкретных задач пользователя. Причем для простых задач обработки ЕЯТ реализу-

ющие их алгоритмы могут войти в состав ЯОИС.

***Выводы***

Рассмотрен общий подход к проблеме анализа ЕЯТ, особенности построения лингвистических моделей и критерии понимания ЕЯТ, в результате чего предложена формальная постановка задачи анализа ЕЯТ, в которой выделены основные подзадачи, связанные с вычислением отношений типизации лексики ЕЯ на лексико-смысловом континууме и интерпретацией некоторого текста на заданной предметной модели. В контексте разработанной архитектуры подсистемы ЯОИС предложена формальная модель обработки ЕЯТ, для которой показано однозначное соответствие процессов обработки естественно-языковой информации и средств (архитектурно-структурных блоков) их реализации.

## СПИСОК ЛИТЕРАТУРЫ

1. *Палагин А.В, Крывый С.Л., Петренко Н.Г.* Знание-ориентированные информационные системы с обработкой естественно-языковых объектов: основы методологии и архитектурно-структурная организация. – УСиМ, 2009. – № 3. – С.42-55.

2. *Боргест Н.М.* Границы онтологии проектирования // Онтология проектирования. – 2017. – Т.7. – № 1(23). – С.7-33. https://doi.org/10.18287/2223-9537-2017-7-1-7-33

3. *Palagin A.V.* An Ontological Conception of Informatization of Scientific Investigations. Cybern Syst Anal 52. – 1-7 (2016). https://doi.org/10.1007/s10559-016-9793-6

4. *Рыков В.В.* Обработка нечисловой информации. Управление знаниями. – М.: МФТИ, 2007. – 156 с.

5. *Демьянков В.З.* Основы теории интерпретации и ее приложения в вычислительной лингвистике. – М.: Изд.-во Моск. ун-та, 1985. – 76 с.

6. *Palagin A.V.* Arrangement and functions of a “language” world picture in semantic interpretation of natural languages and messages // International Journal Information Theories & Application, 2000. – Vol. 7. – № 4 – P.155-164.

7. *Палагин А.В, Крывый С.Л., Петренко Н.Г.* Онтологические методы и средства обработки предметных знаний // [Монография]. – Луганск: изд. ВНУ им. В. Даля, 2012. – 324 с. – Available at: http://www.aduis.com.ua/Monography.pdf.

8. *Петренко М.Г.* Особливості розробки знання-орієнтованого лінгвістичного процесора. – Комп‘ютерні засоби, мережі та системи, 2006. – № 5. – С.18-22.

9. *Леонтьева Н.Н.* К теории автоматического понимания естественных текстов. Часть 2 // Семантические словари: состав, структура, методика создания. – М.: МГУ, 2001. – 41 с.

10. *Палагін О.В., Петренко М.Г.* Архітектурно-онтологічні принципи розбудови інтелектуальних інформаційних систем. – Математичні машини і системи, 2006. – № 4. – С.15-20.

11. *Палагін О.В., Петренко М.Г.* Модель категоріального рівня мовно-онтологічної картини світу. – Математичні машини і системи, 2006. – № 3. – С.91-104.

12. *Палагин А.В., Петренко Н.Г.* К проектированию онтолого-управляемой информационной системы с обработкой естественно-языковых объектов. – Математические машины и системы, 2008. – № 2. – С.14-23.

Поступила в редакцию 05.04.2020

**ДО ПИТАННЯ КОМП’ЮТЕРНОЇ ОБРОБКИ ПРИРОДНОМОВНИХ ТЕКСТІВ**

***Палагін О.В., Петренко М.Г., Зеленцов Д.Г.***

*Розглянуто загальний підхід до проблеми аналізу природномовної інформації, який містить реалізацію низкиу інформаційних технологій, тим чи іншим чином пов’язаних з мовним моделюванням. Крім розробки зазначених інформаційних технологій, необхідно розробити формальну теорію комп’ютерної обробки знань, вилучених з природномовних текстів. Проаналізовано особливості побудови лінгвістичних моделей і критерії розуміння природно-мовних текстів. При цьому виникає низка проблем. Перша проблема зводиться до проблеми аналізу текстової інформації, наданої на природній мові (морфологічний, синтаксичний, семантичний і логічний аналіз) з метою отримання знань. Друга проблема пов’язана з проектуванням системи пошуку, обробки та вилучення знань, розробки і побудови її архітектури, а також інструментарію для користувача. І третьою проблемою є розробка процедур інтеграції знань з кількох предметних областей для забезпечення ефективності здійснення досліджень міждисциплінарного і трансдисциплінарного характеру. Також необхідно приділити особливу увагу питанням використання вже напрацьованих теоретичних положень і практичних рішень. Запропонована формальна постановка задачі аналізу природномовних текстів, в якій виділені основні підзадачі, пов’язані з обчисленням відношень типізації лексики природної мови на лексико-смисловому континуумі та інтерпретацією деякого тексту на заданій предметній моделі. У контексті розробленої архітектури мовно-онтологічної інформаційної системи запропонована формальна модель обробки природно-мовних текстів, для якої показано однозначна відповідність процесів обробки природномовної інформації і засобів (архітектурних блоків) їх реалізації.*

**Ключові слова:** природномовний текст, лінгвістична модель, мовно-онтологічна інформаційна система, аналіз і розуміння природномовних текстів.

## ON THE PROBLEM OF COMPUTER PROCESSING OF NATURAL LANGUAGE TEXTS

*Palagin A.V.*[a], *Petrenko N.G.*[a], *Zelentsov D.G.*[b]

[a] V.M. Glushkov Institute of Cybernetics, Kiev, Ukraine

[b] Ukrainian State University of Chemical Technology, Dnipro, Ukraine

*The present paper deals with the general approach to the problem of analyzing natural language information, including the implementation of a number of information technologies related in one way or another to language modeling. In addition to the development of the aforementioned technologies, it is necessary to develop a formal theory of computer processing of knowledge extracted from natural language texts. The specific features of constructing linguistic models and the criteria for understanding natural language texts are analyzed. This raises a number of problems. The first problem comes down to the problem of analyzing textual information presented in natural language (morphological, syntactic, semantic and logical analysis) in order to extract knowledge. The second problem is associated with designing a system for searching, processing and extracting knowledge, developing and constructing its architecture, as well as tools for the user. The third problem is the development of procedures for the integration of knowledge from several subject areas to ensure the effectiveness of conducting studies of an interdisciplinary and transdisciplinary nature. It is also necessary to pay special attention to the use of already developed theoretical principles and practical solutions. A formal statement of the problem of the analysis of natural language texts is proposed, in which the main subtasks are identified, associated with the calculation of typing relationships of vocabulary of a natural language on a lexico-semantic continuum and the interpretation of some text on a given subject model. In the context of the developed architecture of the linguistic-ontological information system, a formal model for processing natural-language texts is proposed, for which an unambiguous correspondence of the processes of processing natural-language information and means (architectural blocks) of their implementation is shown.*

**Keywords**: ELRE natural language text, linguistic model, language ontological information system, analysis and understanding of natural language texts.

### REFERENCES

1. Palagin A.V, Kryvyj S.L., Petrenko N.G. *Znanie-orientirovannye inphormatsionnye sistemy s obrabotkoy estestvenno-yazykovykh ob'ektov: osnovy metodologii i arhitekturno-strukturnaya organizatsiya* [Knowledge-oriented information systems with the processing of natural language objects: the fundamentals of methodology and architectural and structural organization], USiM, 2009, no. 3, pp.42-55. (in Russian).

2. Borgest N.M. *Granitsy ontologii proektirovaniya* [The boundaries of the design ontology]. *Ontologiya proektirovaniya* [Design ontology], 2017, vol. 7, no. 1(23), pp.7-33. https://doi.org/10.18287/2223-9537-2017-7-1-7-33. (in Russian).

3. Palagin A.V. An Ontological Conception of Informatization of Scientific Investigations. Cybern Syst Anal 52, 2016, 1-7. https://doi.org/10.1007/s10559-016-9793-6

4. Rykov V.V. *Obrabotka nechislovoy inphormatsii. Upravlenie znaniyami* [Processing of non-numeric information. Knowledge management]. M.: MPhTI, 2007, 156 p. (in Russian).

5. Demyankov V.Z. *Osnovy teorii interpretaцii i ee prilozheniya v vychislitelnoy lingvistike* [Foundations of the theory of interpretation and its applications in computational linguistics]. M.: Izd.-vo Mosk. un-ta [Moscow University Publishing House], 1985, 76 p. (in Russian).

6. Palagin A.V. Arrangement and functions of a "language" world picture in semantic interpretation of natural languages and messages. International Journal Information Theories & Application, 2000, vol. 7, no. 4, pp.155-164.

7. Palagin A.V, Kryvyy S.L., Petrenko N.G. *Ontologicheskie metody i sredstva obrabotki predmetnykh znaniy* [Ontological methods and means of processing subject knowledge]. Lugansk: izd. VNU im. V. Dalya [Lugansk: ed. VNU them. V. Dahl], 2012, 324 p. Available at: http://www.aduis.com.ua/monography.pdf. (in Russian).

8. Petrenko M.G. *Osoblyvosti rozrobky znannya-orientovanogo lingvistychnogo protsesora* [Features of the development of a knowledge-based linguistic processor]. Komp'yuterni zasoby, merezhi ta systemy [Computers and systems], 2006, no. 5, pp.18-22. (in Ukrainian).

9. Leont'eva N.N. *K teorii avtomaticheskogo ponimaniya estestvennykh tekstov. Chast 2* [Towards the theory of automatic understanding of natural texts. Part 2]. *Semanticheskie slovari: sostav, struktura, metodika sozdaniya* [Semantic dictionaries: composition, structure, method of creation]. M.: MGU, 2001, 41 p. (in Russian).

10. Palagin O.V., Petrenko M.G. *Arhitekturno-ontologichni pryntsypy rozbudovy intelektualnykh inphormatsiynykh system* [Architectural and ontological principle of development of intellectual information systems] *Matematichni mashyny i systemy* [Mathematical machines and systems], 2006, no. 4, pp.15-20. (in Ukrainian).

11. Palagin O.V., Petrenko M.G. *Model kategorial'nogo rivnya movno-ontologichnoi kartyny svitu* [The model of the categorical level of the movable-ontological picture of the world]. *Matematychni mashyny i systemy* [Mathematical machines and systems], 2006, no. 3, pp.91-104. (in Ukrainian).

12. Palagin A.V., Petrenko N.G. *K proektirovaniyu ontologo-upravlyaemoy inphormatsionnoy sistemy s obrabotkoy estestvenno-yazykovykh ob'ektov* [On the design of an ontologically-controlled information system with the processing of natural language objects]. *Matematicheskie mashiny i sistemy* [Mathematical machines and systems], 2008, no. 2, pp.14-23 (in Russian).