# Особенности лексико-морфологического анализа при извлечении информационных объектов и связей из текстов естественного языка

# Peculiarities of lexical-grammatical analysis for object extraction from natural language texts

**Кузнецов И. П.** (igor-kuz@mtu-net.ru),
**Сомин Н. В.** (somin@post.ru)

Институт проблем информатики РАН, Москва

Анализируется опыт построения семантико-ориентированных лингвистических процессоров, выделяющих структуры знаний из текстов естественного языка (ЕЯ). Одним из важнейших компонент таких систем является блок лексико-морфологического анализа. В процессе развития данного класса процессоров этот блок постоянно совершенствовался и приобрел много новых функций, выходящих за рамки возможностей существующих блоков подобного типа. Данный блок генерирует лексические, морфологические, семантические признаки слов, выявляет простейшие формы естественного языка, имеет специальные средства настройки на предметную область и на особенности текстов ЕЯ. В работе рассматриваются эти функции.

## 1. Введение

### 1.1. Системы с возможностью обучения языку

На протяжении последних 20 лет в ИПИ РАН активно развивается область, связанная с анализом текстов естественного языка (ЕЯ) с целью извлечения полезной информации, формирования структур знаний и их использования для решения прикладных задач — поисковых, логико-аналитических. В 90-х годах был создан класс экспериментальных систем, которые обладали уникальной особенностью — их можно было обучать естественному языку и в дальнейшем использовать для задач фактографического поиска и принятия решений. Это системы ДИЕС [3, 17], СПРУТ, LOG [5], ЭССЕИСТ [6], ИКС[3, 4]. Эти системы создавались в рамках соответствующих пионерных проектов ИПИ РАН, поддерживаемых госбюджетом. Для обучения языку был разработан специальный интерфейс, с помощью которого вводились не только морфологические признаки, но и семантическая компонента каждого нового слова [17]. Для существительных нужно было указать, к какому классу они относятся. Для глаголов — модель управления и т. д. Для ввода новых слов не требовалось каких-либо специальных знаний. В системе ДИЕС это мог делать практически любой грамотный носитель языка — русского (для ИКС — и английского). В результате формировалась база лингвистических знаний слов и словосочетаний (с их семантикой), которая использовалась для лексико-морфологического и синтаксического анализа предложений ЕЯ с их отображением на структуры знаний, составляющие базу предметных знаний (БЗ). Система осуществляла полный разбор предложений с извлечением структур знаний (фактов). При этом учитывались случаи полисемии глаголов, восстанавливалась информация, заданная в неявном виде, и многое другое. Фактографический поиск и другие приложения осуществлялись на уровне структур знаний.

Для создания подобных систем требовались специальные языки представления знаний и инструментальные средства их обработки. Язык — это структурный объект на всех его уровнях, — от поверхностного до семантического. Для обработки

конструкций языка были созданы: язык расширенных семантических сетей (РСС), обеспечивающий представление текстов ЕЯ на уровне структур знаний с любой требуемой точностью, и язык ДЕКЛ для преобразования структур в виде РСС [1, 2, 17].

В тоже время, развитие подобных систем требовало достаточно трудоемкой работы по вводу новых слов. Системы не могли работать с предложениями, где было много незнакомых для них слов. В связи с этим системы типа ДИЕС нашли применение для создания языков экспертных систем, содержащих ограниченное количество слов и форм, достаточных для ввода экспертных знаний. Одна из них — это система СПРУТ, предназначенная для выявления организованных преступных групп

### 1.2. Объектно-ориентированные процессоры

В связи со сказанным в конце 90-х годов в рамках проектов ИПИ РАН начало развиваться другое направление, при котором не требовалось отображения семантики всех предложений на структуры знаний в БЗ [8, 9, 10]. Учитывался тот факт, что определенные категории пользователей интересуются конкретной информацией, которая встречается в текстах ЕЯ. Нужно извлекать из текстов только эту информацию. Данное направление возникло в связи с прикладными разработками для ГУВД г. Москвы. Их проблемы заключались в наличии потоков документов на ЕЯ (сводок происшествий, справок по уголовным делам, обвинительных заключений и др.), в которых было много полезной информации. Это фигуранты, их адреса, телефоны, оружие, автотранспорт и др. Следователей и аналитиков интересовали именно такого сорта объекты и связи между ними. Использование типовых БД требовало громадной работы для их заполнения.

В связи с этим в ИПИ РАН была инициирована работа по созданию лингвистических процессоров (ЛП), обеспечивающих автоматическое выделение их текстов ЕЯ информационных объектов и связей с формированием структур знаний в БЗ. Такие ЛП были названы ***объектно-ориентированными*** (в некоторых работах — ***семантико-ориентированными***). Для успешного создания подобных ЛП у разработчиков уже имелась достаточная база. В результате была создана система «Криминал», обеспечивающая автоматическое извлечение структур знаний из текстов ЕЯ и их использование для решения логико-аналитических задач — для следователей и аналитиков. В данной системе не требовалось вводить морфологические и другие характеристики слов. Для этого был создан блок ***лексико-морфологического анализа*** (ЛМА), который анализирует текст и строит семантическую сеть (РСС), названной ***пространственной структурой текста*** (ПС-текста). Последняя обрабатывается блоком ***синтактико-семантического анализа*** (ССА), который (на языке ДЕКЛ) анализирует ПС-текста и формирует на РСС структуру, представляющую объекты и связи между ними. Такие структуры образуют БЗ.

Отметим, что блок ЛМА написан на языке Си++, при использовании которого на определенных этапах формализации текстов возникают существенные трудности. В тоже время, чем больше функций берет на себя блок ЛМА, тем в большей степени снимает трудности дальнейшего процесса формализации, который осуществляется блоком ССА [13, 14, 20].

Поэтому в последующих проектах ИПИ РАН (АНАЛИТИК, ПОТОК и др.) блок ЛМА постоянно усовершенствовался [7, 18]. В процессе выполнения проектов объектно-ориентированный ЛП использовался в различных предметных областях для формализации различных корпусов текстов. В связи с этим в блок ЛМА постоянно вводились новые возможности. В данной статье обобщается опыт разработчиков по построению объектно-ориентированных ЛП. Статья посвящена описанию особенностей блока ЛМА, выполняющего сложные функции по анализу текстов ЕЯ и обеспечивающего необходимой информацией блок ССА для данного типа ЛП.

## 2. Особенности объектно-ориентированных ЛП

Наш опыт показывает, что при наличии потока документов, требующих обработки, учесть все формы и особенности ЕЯ, используемые при описании многих объектов и связей, и построить сколь либо полную «модель языка» — неразрешимая задача. Поэтому требуется постоянное совершенствование ЛП.

В связи с этим рамках проектов ИПИ РАН развивается направление, когда программа объектно-ориентированного ЛП отделяется от ***лингвистических знаний*** (ЛЗ). Последние определяют всю процедуру анализа (см. ниже). ЛЗ имеют вид декларативных структур, которые легко менять и настраивать. В нашем случае роль таких структур выполняют фрагменты РСС [12–14]. Настройка ЛП осуществляется только за счет разработки ЛЗ.

Задача ЛП — поддерживать ЛЗ. При использовании подобных ЛП облегчается настройка на корпуса текстов, особенности предметной области. Корректировать ЛЗ может человек, обученный формализму РСС и знакомый с элементами математической лингвистики. Ему не нужно уметь программировать. Тогда возникает вариант, когда один человек может настраивать лингвистический процессор — находить ошибки и устранять их.

Как говорилось ранее, объектно-ориентированный ЛП состоит из блоков лексико-морфологического и синтактико-семантического анализа.

Блок **лексико-морфологического анализа** (ЛМА), выделяет из документа слова и предложения и выдает в виде семантической сети (ПС-документа), представляющей последовательность компонент (слов в нормальной форме, чисел, знаков) и их основные признаки. Блок ЛМА имеет три основных подсистемы:

- Лексический анализатор, который ответственен за правильное деление входного текстового потока на абзацы, предложения и слова (формирует лексические признаки слов);
- Морфологический анализатор, осуществляющий морфологический анализ всех слов текста (приводит слова в нормальную форму и формирует для них морфологические признаки);
- Систему предметных словарей, призванную распознать в тексте характерные термины (формирует семантические признаки).

Блок ЛМА имеет свои лингвистические знания (ЛЗ) — средства ***параметрической настройки***, позволяющие учитывать разнообразие текстовой типологии, и набор ***предметных словарей*** (словарь стран, регионов России, имен, профессий и др.) для придания словам и словосочетаниям дополнительных семантических признаков [13, 18].

Блок **синтактико-семантического анализа** (ССА) путем анализа ПС-документа выделяет объекты и связи. На их основе строит другую семантическую сеть, представляющую ***семантическую структуру документа*** (СС-документа), называемую также ***содержательным портретом*** [12, 15, 19]. Этот блок включает в себя ***базу лингвистических знаний*** (ЛЗ), которая содержит правила анализа текста во внутреннем представлении (РСС). Они определяют работу ЛП [13–19].

Блок ССА управляется ЛЗ, за счёт которых обеспечивается:

- извлечение информационных объектов (лиц, организаций, событий, их места);
- выявление связей объектов; например, связей лиц с организациями, адресами и др.;
- анализ глагольных форм, причастных и деепричастных оборотов с выявлением фактов участия объектов в тех или иных действиях;
- идентификация объектов с учетом анафорических ссылок и сокращенных наименований;
- выявление связей действий с их местом или временем (где и когда имело данное действие или событие);
- анализ причинно-следственных и временных связей между действиями и событиями.

Особенности блока ССА описаны во многих статьях, в том числе трудах конференций Диалог [8–16]. Гораздо меньше внимания уделялось описанию работы блока ЛМА. В данной статье будет восполнен этот пробел.

Блок ЛМА [7], разработанный для русского и английского языков, основан на традиционной для таких блоков схеме словарей. Однако, помимо этого, в блоке ЛМА присутствует еще ***словарь обобщенных основ***, позволяющая обрабатывать и новые слова, см. п. 5.

Блок ЛМА приводит слова в нормальную форму и присваивает им признаки, которые делятся на три группы: — лексические признаки (слово с большой буквы, большими буквами, с точкой на конце или это отдельная буква и др.)

- морфологические признаки (грамматическая категория слова, число для существительных и т. д.);
- семантические признаки (имя, организация, оружие и др., а также ключевые слова, относящиеся к соответствующему типу объектов).

Предусмотренный лексикографический анализ обеспечивает автоматическое деление текста на самостоятельные части (например, выделение документов из сводок) и определение начала и конца предложения, а также начала и конца абзаца.

Выходная информация бока ЛМА (т. е. ПС-текста) сохраняет порядок предложений в тексте, разделяя их фрагментами типа SENT, и порядок слов в предложении. При этом каждое слово представляется с его признаками, см. п. 7.

## 3. Предметные области и тексты

В настоящее время имеется большой опыт использования объектно-ориентированных ЛП в различных предметных областях, где требуется выделение различных объектов из корпусов текстов со своими особенностями. В данном разделе мы постараемся обобщить эти особенности и связанные с ними трудности, которые требовали постоянного совершенствования блока ЛМА. С какими предметными областями и текстами мы имели дело:

### 3.1. Документы криминальной милиции

Работа делалась по заказу ГУВД г. Москвы [9]. Была создана система «Криминал», в БЗ которой были введены: сводки происшествий по (более 500 тыс. происшествий), справки по уголовным делам (несколько сотен), обвинительные заключения (около сотни), записные книжки фигурантов (около сотни). Система обеспечивает выделение фигурантов, их примет, связей, организаций, дат, документов, номеров счетов, оружия (всего до 40 типов объектов) с указанием характера их участия в криминальных действиях .

### 3.2. Резюме (для приема на работу) на русском и английском языках

Работа инициировалась компанией HEADHUNTER и имела целью автоматическую обработку архивов произвольно написанных резюме и их представление в формате сайта данной компании [16]. Была создана система, выделяющая из резюме атрибуты человека, места его работы, учебы, соответствующие периоды времени, знание языков и т. д. Система отлаживалась в несколько этапов. Вначале на выборках в различных областях (информационные технологии, банковское дело, финансы, юриспруденция и др.) по 200 резюме. Далее отладка шла на специально подобранных «критичных» резюме (которые при обработке давали шумы по тем или иным типам объектов), из которых составлялись специальные выборки. Система работала на сайте упомянутой компании, чтобы автоматически переводить резюме пользователей, поступающих через Интернет, в формат сайта.

### 3.3. Документы о терроризме на русском языке

Работа носила инициативный характер с целью внедрения в крупный проект. Система дополнительно выделяла руководящих лиц, правительственные организации, террористов (как свойство фигурантов), террористические организации, орудия преступления, время и место событий и т. д., а также связи и участие лиц в тех или иных действиях. На первом этапе она отлаживалась на массиве в 300 документов, относящихся к террористической деятельности. В дальнейшем отладка шла на материалах СМИ, в том числе, взятых из Интернет [18]. Была также разработана ДЕМО-версия для обработки документов на английском языке.

### 3.4. Документы о памятниках культуры

Работа носила инициативный характер — делалась для Министерства культуры. Система выделяет из текстов тип памятника (скульптура, монумент), кто является автором, создателем, время, место и многое другое.

Во всех случаях (за счет средств настройки блоков ЛМА и ССА) удавалось добиться требуемого качества работы ЛП [10, 13, 17].

Отметим высокое разнообразие перечисленных предметных областей, которое определяется не только в различием выделяемых объектов и связей. Еще большие отличия можно наблюдать в «стиле» текстовых сообщений, связанных с предметными областями. В понятие «***стиль***» мы включаем весь комплекс особенностей, присущих определенной группе текстов. Сюда входят:

- лексика предметной области, включая всю совокупность специфических терминов предметной области;
- коммуникативный тип текста: художественное произведение, техническая или аналитическая статья, новостное сообщение, приказ, PR- текст (например — реклама);
- структурный тип текста: связный текст, список, таблица, математическая формула;
- инструмент создания текста (имеется в виду текстовый редактор или генератор текста, с помощью которого получен текст);
- способ грамматического оформления текста, под которым понимается следование стандартным правилам орфографии языка (проставление необходимых знаков препинания и разделителей, позволяющих структурировать текст);
- следование принятой в языке орфографии, что выражается в количестве орфографических ошибок или нарочитого введения искаженной лексики.

Отметим, что резкое увеличение разнообразия текстовой типологии, с которой мы столкнулись в различных предметных областях. В значительной степени это вызвано бурным распространением Интернет и тем фактом, что порождение текстов все в большей мере стали осуществлять люди различной степени подготовки и грамотности. Как следствие — наличие значительного количества специальных разделителей, отсутствие знаков препинания, большое количество сокращений, ошибок и многое другое. Отсюда следуют дополнительные требования к компонентам блока ЛМА и средствам их настройки. Рассмотрим их подробнее.

## 4. Лексический анализатор

Эта компонента блока ЛМА имеет дело с целым рядом взаимосвязанных задач, решение которых совершенно необходимо для успешной работы всего ЛП. Рассмотрим их особенности.

Прежде всего, решается **задача структуризации текста**. Дело в том, что текст в современной информационной среде — сложно структурированный объект. И его структура должна быть распознана и аккуратно передана блоку ССА, поскольку от правильного распознавания структуры текста в значительной степени зависит корректность всего анализа. Поэтому задача структуризации распадается на цепочку локальных задач.

### 4.1. Задача выделения лексем

При выделение из входного потока лексем (слов, знаков препинания, разного рода раздели-

телей и др.) имеют место следующие трудности. Современный деловой текст содержит большое количество лексем, являющихся техническими, административными и фирменными названиями, телефонами, шифрами, номерами автомобилей, адресами электронной почты и Интернет и проч., содержащими цифры, буквы и разделители практически в произвольной комбинации. Такие знаки, как «-», «.» и «,», доставляют много хлопот при их анализе, в одних случаях являясь разделителями лексем, а в других — нет.

### 4.2. Задача выделения предложений

В виду огромного разнообразия текстовых «стилей», по отношению к современным текстам становится трудно говорить о предложении. Скорее следует говорить о «сильносвязанных» отрезках текста, в которых идет речь об одном объекте или одной ситуации, в которой участвуют несколько взаимодействующих объектов. В результате само понятие «предложение» резко расширяется, включая в себя, помимо обычных предложений (с точкой в конце), еще массу различных текстовых отрывков: ячеек таблицы, элементов списка и прочих, грамматическое оформление которых нетрадиционно.

### 4.3. Задача выделения абзацев

Абзацем мы называем отрезок текста из одного или нескольких предложений, связанных единой темой. Опять-таки расплывчатость этого определения позволяет трактовать его широко. Однако для блока ССА понятие абзаца является весьма важным, поскольку многие его механизмы направлены именно на идентификацию и совмещение объектов внутри одной темы. Лексический анализатор содержит в своем составе ряд алгоритмов, выделяющих абзацы, причем — разных типов.

Надо сказать, что задачи выделения предложений и абзацев весьма нетривиальны. Трудности выделения абзацев главным образом связаны с тем, что хорошо различимые разделители абзаца — пустые строки, отступы, границы клеток таблицы — теряются или искажаются при преобразовании текстов. Но гораздо большие трудности возникают при идентификации предложений. Дело в том, что современные пользователи Интернет вообще не считают необходимым ставить точки в конце предложения. В тоже время точка активно используется в качестве ограничителя сокращений, разделителя между частями несловарного компонента (электронного адреса, многозначного числа, банковского номера и пр.). Кроме того, разделителем предложения может являться не только точка, но и другие знаки («;», «:», «!», «?», «|» и т. д.). В результате задача разбиения текста на предложения становится просто головоломной шарадой, требующей учета массы разного рода частных правил и исключений.

### 4.4. Задачи унификации текста

Естественный язык — система необычайно многовариантная: один и тот же смысл (приблизительно) может быть выражен умопомрачительным количеством его текстовых выражений. Задача лексического анализатора: унифицировать написание отдельных слов и сокращений, привести к стандартной форме написание ряда стандартных словосочетаний. Трудность тут в том, чтобы выявить наиболее употребительные лексемы и словосочетания, требующие унификации.

К этой сложности примыкает проблема обнаружения и (по возможности) исправления ***опечаток и грамматических ошибок***. В современных текстах их — громадное количество, и бороться с ними — задача из сложнейших. Кроме того, в современных текстах, особенно из Интернет, намечается тенденция нарочитого переделывания и перевирания слов, типа «*ацкий ужос*» или «*падстол*». Начинает формироваться целая интернетная «феня». В связи с этим потребуется постоянная корректировка языковых словарей и правил составления предложений.

Еще одна важная функция лексического анализатора — определение ***лексических признаков*** слов. Примеры такого рода признаков: «Слово из кириллицы с прописной буквы». «слово из кириллицы из прописных букв», «разделитель», «слово из латинских букв» и проч., всего — около 20 лексических признаков. Лексические типы являются важной дополнительной информацией, облегчающей работу как морфологического анализатора, так и блока ССА.

Наконец, лексический анализатор для ряда слов способен выполнить семантический анализ, определяя по формальному виду слове его ***семантическую категорию***. К этому случаю относятся, скажем, сокращения имен и отчеств: прописная буква, за которой идет «.». Например «*А.*», «*Н.*», «*J.*». Еще примеры идентифицируемых семантических классов: «адрес электронной почты», «Интернет-адрес» (URL), «целое число», «число с дробной частью». Собственно, определение семантического класса каждого слова или словосочетания является одной из задач всего ЛП. И чем раньше такой класс будет определен, тем легче дальнейший анализ.

## 5. Морфологический анализатор

Задача морфологического анализатора — нормализация слов, определение морфологических признаков лексем, а также (в ряде случаев) — на-

хождение его семантического класса. Отметим, что к настоящему времени разработан целый ряд морфологических анализаторов русского языка, среди которых упомянем лишь некоторые: [21, 22, 23, 24].

### 5.1. Базовая схема анализа

Первоначально была реализована базовая схема анализа [5]. Считается, что каждое слово имеет постоянную часть (основу) и переменную часть. Последняя образует словоизменительную парадигму или класс окончаний. Были накоплены два словаря: словарь классов окончаний (СКО), в котором хранятся все возможные парадигмы русского языка и словарь основ (СО), в котором хранятся основы слов со ссылками на соответствующий класс окончаний.

Например, слово «*бытие*» имеет основу «*быти*» и класс окончаний за номером 1759, содержащий окончания в именительном, родительном, дательном, винительном, творительном и предложном падежах, а именно: «*е*», «*я*», «*ю*», «*е*», «*ем*», «*и*» (множественного числа это слово не имеет). Соответственно в СО имеется запись «*быти 1759*», а в СКО под номером 1759 закодирована парадигма с указанными окончаниями.

Отметим, что в общем случае в СО может быть несколько записей с одинаковой основой (но с разными классами окончаний), а на один и тот же класс окончаний может ссылаться несколько слов с разными основами. Возможны случаи пустой основы (пример: «*хорошо*»-«*лучше*») и пустого класса окончаний (для неизменяемых слов). Кроме основы и вариантов окончаний, в СКО хранятся морфологические признаки, соответствующие определенному классу окончаний в целом (постоянная морфологическая информация) и каждому окончанию парадигмы в отдельности (переменная морфологическая информация). Так, для класса 1759 в качестве постоянной информации хранятся признаки существительного, среднего рода, неодушевленности и второго склонения, а для каждого окончания хранится признак соответствующего падежа.

Алгоритм морфологического анализа при наличии данных словарей сводится к следующему. Для слова рассматриваются все варианты его разбиения на основу и окончание. Если для данного варианта разбиения находится основа, а в соответствующем ей классе окончаний находится вариант окончания, то данный морфологический разбор является корректным и слово получает морфологические признаки, взятые из постоянной и переменной частей морфологической информации. В общем случае может быть найдено и выдано несколько вариантов морфологического разбора, что известно, как морфологическая омонимия.

### 5.2. Морфологический анализ незнакомых слов

В принципе предложенная схема анализа вполне корректна. Однако на практике ее успешное использование достаточно проблематично. Дело в том, что такая схема предполагает ручную разработку обоих словарей. И заметим — не только первоначальную разработку, но и их постоянное пополнение. Последнее обстоятельство особенно неприятно: в русском языке — более 100 тыс. слов общеупотребительного назначения и миллионы специальных терминов. Кроме того, перестройка вылилась в активную языковую экспансию: в русскоязычных текстах стало использоваться огромное количество англоязычных слов, которые никогда не входили в классические словари русского языка. В результате при обработке таких текстов система «натыкалась» на множество слов, отсутствующих в СО. Фактически требовалось ежедневное пополнение словаря. Но в то же время, создание таких словарей требует высокой лингвистической квалификации и исключительной тщательности. Составление достаточно полных морфологических словарей — кропотливая работа, требующая десятилетий.

Выход из описанной ситуации известен — обработка незнакомых системе слов «по аналогии» [24, 25]. В нашей реализации этого метода использовался третий словарь — «***словарь хвостов основ***» (СХО). В словарь записываются все 1-буквенные, 2-буквенные, 3-буквенные и т. д. «хвосты» основ (первые буквы основ отбрасываются) с указанием соответствующего класса окончаний. Было решено, что в СХО не будет одинаковых «хвостов», а его класс окончаний вычисляется из статистических соображений — по максимуму основ в СО, имеющих данный «хвост» и данный класс окончаний. Если слово не находится в словаре СО, то та же схема анализа повторяется, но уже с помощью пары словарей СХО-СКО.

В реализации словари СО и СХО были слиты в один словарь, за которым закрепилось название обобщенного словаря основ (ОСО), в результате чего все варианты анализа, — как точные, так и по аналогии, — выявляются за один проход по словарю. Кроме того, был разработан способ сжатия словарной информации, который позволил хранить все словари в оперативной памяти существующих на момент реализации (1996 г.) компьютеров (объем словаря на 90 тыс. основ составляет 894 КБ).

### 5.3. Борьба с морфологической омонимией

Ясно, что использование обобщенного словаря основ ОСО может приводить к лишним вариантам морфологического анализа. Было предложено два достаточно эффективных способа борьбы с морфологической омонимией.

Первый способ — эмпирический алгоритм, отбрасывающий наименее вероятные варианты морфологического анализа. Такая «зачистка» вариантов выполняется по многим критериям, учитывающим наличие слова в СО, длину основы с СХО, часть речи. Кроме того, эмпирический алгоритм расставляет все варианты разбора в порядке их вероятности. Такое ранжирование необходимо для ряда приложений, когда используется только один вариант морфологического анализа.

Второй способ — частичный синтаксический анализ. Дело в том, что в предложении слово вступает в синтаксические связи с другими словами, и выявление этих связей позволяет отбросить варианты морфологического анализа, этим связям не удовлетворяющих. Прежде всего было реализовано распознавание двух конструкций: полного согласования и генетической цепочки.

### 5.4. Особенности блока английской морфологии

Помимо русского морфологического словаря был создан и английский морфологический словарь. Он использует уже разработанное для русской морфологии программное обеспечение, которое оказалось возможным адаптировать к специфике английского языка. Блоки английской и русской морфологии выдают практически одни и те же морфологические характеристики. Это позволяет использовать для синтаксико-семантического анализа англоязычных текстов те же средства, что и для русского языка. В результате появилась возможность записывать лингвистические знания для этих языков в одном и том же формализме.

Общий объем словаря основ блока английской морфологии — около 85 тыс. Тем не менее, для повышения качества работы этого блока в него был введен ряд специфических для английского языка алгоритмов, которые в основном касаются отсева лишних вариантов морфологического анализа. Дело в том, что слова английского языка чрезвычайно омонимичны. Очень часто одно и то же слово может быть и существительным, и глаголом, и прилагательным. В блоке английской морфологии реализованы алгоритмы, позволяющие в ряде случаев корректно отбрасывать лишние варианты (другие варианты отсеиваются в процессе синтактико-семантического анализа). Блок ЛМА был модифицирован для работы с предметными словарями английского языка, которые удалось совместить со словарями русского языка.

## 6. Система предметных словарей

Предметные словари (стран, имен собственных, организаций, профессий, видов оружия и др.) состоят из терминов. Множество словарей образует систему.

***Система предметных словарей*** (СПС) предназначена для распознавания в тексте слов и словосочетаний, специфичных для конкретной предметной области. Им присваиваются признаки принадлежности к определенной семантической категории. Будем называть этот процесс идентификацией терминов словаря. Такая принадлежность является основой выделения объекта. В предметном словаре может быть или термин, представляющий объект определенного типа (но таких объектов может быть достаточно много), или характеристическое слово, опираясь на которое можно начинать распознавание объекта — на уровне синтактико-семантического анализа.

Видимо, без СПС не обходится ни один серьезный проект лингвистического процессора. В нашей разработке СПС встроена в блок ЛМА. Причина этого — главным образом в быстродействии. Поиск в СПС предполагает частые обращения к ней, а потому требуется высокая эффективность поиска, чего трудно достичь без использования универсальных языков программирования. В нашем случае программное обеспечение СПС написано на Си++.

Структурно СПС состоит из произвольного количества ***словарей***, являющих определенный семантический класс. В каждом из словарей может содержаться произвольное количество ***словарных записей***. Под записью в тривиальном случае понимается термин (однословный или многословный). Однако простыми терминами словарные объекты не ограничиваются — там могут содержаться ***словарные шаблоны***, описывающие группу терминов. Возможности описания словарных шаблонов будут приведены ниже. В настоящее время разработаны более 20 предметных словарей; среди них: «Улицы г. Москвы», «террористические организации», «оружие», «известные личности».

### 6.1. Требования к предметным словарям

К СПС, помимо эффективности, предъявляются еще ряд требований, важнейшими из которых являются:

1) *Требование множественности*. Информационная система может иметь несколько предметных словарей различного содержания, причем число словарей заранее не ограничивается и может динамически пополняться. Общее число словарей может превышать несколько сотен.
2) *Требование к объему предметного словаря.* Каждый словарь, разумеется, содержит множество записей, список которых может расширяться. В нашей постановке задачи предполагается, что объем словаря не может быть заранее ограничен каким-либо фиксированным числом, а должен быть потенциально неограниченным и определяться только

вычислительными мощностями и объемом оперативной памяти и накопителей.

3) *Требование к подготовке информации*. Подготовка текстового материала для загрузки словаря, естественно, выполняется специалистами в соответствующих предметных областях. Поэтому форма исходного вида словаря должна быть максимально простой.
4) *Требование вариативности поиска*. Должна быть предусмотрена корректная обработка случаев, когда написание термина в тексте так или иначе не соответствует каноническому виду термина в словаре.

Если первые три требования носят в основном технический характер, то удовлетворить требованию вариативности поиска в условиях естественного языка — задача весьма непростая. Дело не только в том, что термин может стоять в любом из падежей, что не дает возможности напрямую совместить текст с предметным словарем (эта проблема решается с помощью морфологического анализатора). Основная трудность в другом: в социуме имеет хождение множество вариантов употребления одного и того же термина, и указать их все для разработчика предметного словаря является непосильной задачей. Вот примеры.

Как правило, названия улиц записаны в именительном падеже. Например, «*проживает по адресу Б. Академическая ул. д. 6–18*». Иногда встречается дательный падеж: «*по Б. Академической*». Гораздо более усложняет дело вариативность сокращений и перестановки слов. Например, канонический вид названия одной из улиц Москвы — «*Шипковский 1-й пер.*». Однако, встречаются в текстах написания: «*1-й Шипковский пер.*», «*1-ый Шипковский переулок*», «*п-к 1-вый Шипковский*» и другие варианты. Отметим, что возможна не только перестановка и вариативное написание слов, но и выпадение или добавление слов. Например, «*Туполева Академика наб.*» может быть названа как «*набережная Туполева*», а «*Тихий туп.*» иногда добавляют пояснение «*ул. Тихий туп.*». Кроме того, некоторые сокращения, применяемые авторами текстов, далеко не однозначны. Например «*С.*» может означать «*Северный*» или «*Старый*»; «*Б.*» может означать «*Большой*», а может быть сокращением имени, например «*ул. Б. Галушкина*».

## 6.2. Возможности предметных словарей

Подключение новых словарей может значительно усилить ЛП в плане выделения объектов. Однако для того чтобы словари в самом деле стали действенным и удобным механизмом, необходимо, чтобы они обладали рядом нетривиальных возможностей.

В нашей версии СПС реализованы несколько таких возможностей.

Во-первых, идентификация термина в любом числе и падеже. Например, если в словаре есть термин «*программный продукт*», то в тексте будут распознаваться и соответствующим образом идентифицироваться термины «*программного продукта*», «*программных продуктов*» и т. д. Распознавание выполняет программное обеспечение системы предметных словарей, использующее блок морфологического анализа.

Во-вторых, допускается несколько вариантов написания одного и того же термина. Дело в том, что средствах СМИ и многих других текстах пользуются различными вариантами именования одного и того же объекта, в том числе сокращенным описанием. Например, если в тексте встретилось *Путин, Меркель, президент Франции* и т. д., то понятно, о ком идет речь. Для приведения таких словосочетаний к стандартному виду в словари введена специальная запись. Например, в словаре ФИО может иметь место запись:

*Меркель Ангела*
*= Ангела Меркель*
*= А. Меркель*
*= Меркель*

В данном примере основной термин — «*Меркель Ангела*». К нему будут приводиться все остальные написания этого имени, записанные после символа «=». Эта возможность особенно эффективна при выявлении не только ФИО известных деятелей, но и названий организаций (включая их сокращения), географических названий и др. При этом блок ССА осуществляет дополнительную фильтрацию, например, когда в тексте несколько лиц с фамилией *Меркель* или рядом со словом *Меркель* стоит какое-либо имя, не представленное в предметном словаре.

В-третьих, в предметные словари введена возможность описания группы терминов, у которых лишь первое слово фиксировано, а остальные могут быть описаны с помощью совокупности признаков (лексических и морфологических). Реализованы, так называемые, ***словарные шаблоны***. Например, в словаре допустима строка:

*заведующий* {NOUN,КЕМ}

Такая запись в словаре профессий означает, что подходящими под этот шаблон терминами могут быть все словосочетания, начинающиеся со слова «*заведующий*», за которым идёт существительное (NOUN) в творительном падеже (КЕМ): «*заведующий складом*», «*заведующий библиотеками*» и т. д. Кроме того, в качестве шаблона можно употреблять имя другого (или того же самого) словаря. Это дает возможность точнее указывать те варианты, которые допускает шаблон. Фактически на словари возлагаются элементы синтаксического анализа, позволяющие значительно уменьшить количество записей в словаре, а также облегчить работу блока ССА.

В-четвертых, имеется возможность управлять лексическим и морфологическим анализами в процессе распознавания терминов словарей. Так, например, в словаре террористических организаций может быть указано:

*Организация эта*
*= ЭТА\!*

Это означает, что, благодаря признаку «\», слово «*эта*» в процессе идентификации морфологическому анализу не подвергается (т. е. его каноническая форма совпадает с написанием). И, кроме того, благодаря признаку «!» идентификация совершается, если в тексте слово «*ЭТА*» записано прописными буквами. Эти возможности позволяют повысить точность распознавания, отсеивая ложные вхождения.

Отметим, что язык записи терминов в словарях чрезвычайно прост. Термин пишется в своей канонической форме на отдельной строке (включая, разумеется, указанные выше. дополнительные возможности). Поэтому ввод новых терминов или даже создание новых словарей может быть выполнено пользователем или оператором-лингвистом, не знакомым с особенностями работы ЛП.

Помимо указанных возможностей имеется еще ряд специальных операторов настройки, позволяющих управлять идентификацией терминов для тех или иных словарей.

## 7. Пространственные структуры

Текст ЕЯ — это сложный структурный объект, который в процессе его формализации проходит множество уровней преобразования. На первом уровне работает блок ЛМА, который формирует РСС, называемую ***пространственной структурой текста*** (ПС-текста). Далее следуют преобразования, осуществляемые блоком ССА, которые приводят к формированию ***семантической структуры*** (СС-текста) для БЗ.

Рассмотрим особенности ПС-текста. Информация об абзацах и предложениях представляется в виде фрагмента SENT, с помощью которого представляется:

- позиция первого слова предложения относительно начала входного потока;
- признак начала абзаца и количество разделительных строк;
- номер строки, на которой расположено первое слово предложения.

Для каждого слова (и для каждого варианта его разбора) блок выдает фрагменты типа LR, задающих последовательность слов. В каждом из фрагментов представлено:: нормализованное слово и его порядковый номер. Далее следуют его признаки. Вот некоторые из них: NAME0 — слово начинается с прописной буквы, HEAD_ — слово полностью состоит из прописных букв, NAME1 — инициалы, POINT — пункт, HEAD_1 — слово с прописной буквой, NUM) — целое число, NUM_F — число с дробной частью, ENGL — слово из букв латинского алфавита, WEB_C — URL (адрес Интернет), MAIL_E — адрес электронной почты, FIRST_ — признак первого слова на новой строке, LETT — слово из одной буквы и т. д. (морфологические и семантические признаки).

Фрагменты типа LR и SENT вместе с выделенными признаками — это семантическая сеть (РСС), которая в дальнейшем проходит множество уровней преобразования, осуществляемое блоком ССА.

В общем случае блок ЛМА выдает несколько вариантов разбора. Эта ситуация является весьма типичной. Например, слово «*стекло*» является и существительным и глаголом. Тогда в ПС-текста, помимо фрагмента LR для первого варианта разбора, генерируются фрагменты LD (с их признаками) для других вариантов. Отсев вариантов осуществляется блоком ССА в процессе обработки ПС-текста и построения семантической структуры [12, 14].

## 8. Средства параметрической настройки

Опираясь на опыт построения ЛП для различных предметных областей (см. п. 3), чтобы постоянно учитывать все новые особенности текстовой типологии, в блок ЛМА были введены средства управления лексико-морфологическим анализом, названые средствами ***параметрической настройки***. Эти средства относятся к ЛЗ и размещаются в отдельном файле. Они имеют вид списков, оформленных в виде фрагментов РСС со своими именами. Имена играют роль операторов и определяют вид анализа.

Всего реализовано 27 типов фрагментов. Из них 18 относится к блоку лексического анализа, 5 — блоку морфологического анализа и 4 — предметным словарям.

Лексическое оформление текста — один из самых вариативных аспектов, сильно меняющихся от задачи к задаче. В связи с этим аппарат лексической настройки потребовал значительного развития в плане разработки новых операторов (заданных в виде фрагментов). Рассмотрим их, разделив операторы на смысловые группы.

### 8.1. Средства идентификации начала и конца предложения

- Если слово, указанное во фрагменте NEW_ SENT, записано в тексте с прописной буквы и находится в начале строки, то оно рассматривается как начало нового предложения.

- Если в тексте встречается одно из слов (символов, знаков), указанных во фрагменте END_SENT, то оно считается концом предложения.
- Фрагмент ABBR задает список сокращений с точками на конце, которые считаются цельными словами и точки не рассматриваются как конец предложения
- Фрагмент SEPARATOR задает символы, которые всегда являются разделителями слов.

### 8.2. Средства для замены или удаления некорректных символов или слов

- Фрагменты LETTER_CH и WORD_BAD задают замены (или удаление) нежелательных слов или знаков в тексте.
- Фрагменты BEG_SYMB задают набор удаляемых знаков в начале слова, а END_SYMB — в конце.

### 8.3. Средства унификации и синонимичных замен

- Фрагмент SYNON задает список синонимичных слов, которые заменяются на слово из первой позиции.
- Фрагмент TERMIN_ заменяет слова, записанные на второй и последующих позициях, на слово в первой позиции.
- Фрагмент SIGN_MANY задает повторяющиеся символы, следующие один за другим (например, набор черточек) на один символ (черточку).

### 8.4. Средства настройки морфологического анализатора

- Фрагмент MORF определяет генерацию морфологических признаков слова в виде фрагментов ПС-текста.
- Фрагмент NOMO задает список слов, для которых устанавливается запрет на нормализацию и морфологический анализ.
- Фрагмент NOMOE задает для слов дополнительные признаки, которые вставляются в ПС-текста.

Это необходимый набор операторов, без которых (как оказалось) трудно обеспечить качественный лексико-морфологический анализ многих текстов ЕЯ, и следовательно, качественную работу всего объектно-ориентированного ЛП.

## Заключение

В данной статье рассмотрены особенности блока лексико-морфологического анализа, используемого в объектно-ориентированных лингвистических процессорах (ЛП) при формализации текстов ЕЯ, т. е. для извлечения из них информационных объектов, признаков и связей. Блок обладает уникальными возможностями, с помощью которых обеспечивается устойчивая и качественная работа ЛП при обработке массивов документов на ЕЯ в различных предметных областях: «Криминалистика», «Резюме», «Терроризм», «Памятники культуры» и др.

## Литература


1. *Кузнецов И. П.* Семантические представления// М.: Наука. 1986 г. 290 с.
2. *Кузнецов И. П., Шарнин М. М.* Продукционный язык программирования ДЕКЛ. Сб. Система обработки декларативных структур знаний Деклар-2// ИПИ РАН, 1988.
3. *Кузнецов И. П., Шарнин М. М.* Интеллектуальный редактор знаний на основе расширенных семантических сетей// Сб. Системы и средства информатики. Вып. 5. М. Наука, 1993.
4. *Кузнецов И. П.* Гипертекстовые технологии на семантической основе// Сб. Системы и средства информатики. Вып. 7. М. Наука, 1995.
5. *Любушкина Л. А., Михеев А. С., Соловьева Н. С., Сомин Н. В., Фрейдлин И. Я.* LOG — программа, ведущая диалог на естественном языке // Вторая всесоюзная конференция по ИИ «ВКИИ-90». Минск: Центрпрограммсистем, 1990.
6. *Карунин А. Б., Соловьева Н. С., Сомин Н. В.* ЭССЕИСТ — программа, ведущая диалог с базой знаний на естественном языке. // В кн.: Социальная информатика-93 / Сб. научн. трудов под ред. Колина К. К. и Суслакова Б. А. — М., 1993. — С. 168–174
7. *Сомин Н. В., Соловьева Н. С., Шарнин М. М.* Система морфологического анализа: опыт эксплуатации и модификации// Системы и средства информатики, Вып. 15, 2005, стр. 20–30.
8. *Кузнецов В. П.* Автоматическое выявление из документов значимой информации с помощью шаблонных слов и контекста// Труды межд. Семинара Диалог 98? Т. 2. Казань: ООО «Хетер» 1998.
9. *Кузнецов И. П.* Методы обработки сводок с выделением особенностей фигурантов и происшествий// Труды межд. Семинара Диалог 99. Т. 2. Тарусса, 1999.
10. *Кузнецов И. П., Кузнецов В. П., Мацкевич А. Г.* Система выявления из документов значимой информации на основе лингвистических знаний в форме семантических сетей // Труды межд. Семинара Диалог 2000. Т. 2. Протвино, 2000.
11. *Кузнецов И. П.* Лингвистический процессор для автоматического выявления из текстов значимой информации с ее компоновкой в рамках указанных шаблонов// Труды международного семинара Диалог-2001. Том 2. Протвино, Наука, 2001.
12. *Kuznetsov I., Matskevich A.* System for Extracting Semantic Information from Natural Language Text // Труды международного семинара Диалог-2002 по компьютерной лингвистике и ее приложениям. Т. 2. Протвино, Наука, 2002.
13. *Кузнецов И. П., Мацкевич А. Г.* Особенности организации базы предметных и лингвистических знаний в системе АНАЛИТИК // Труды конференции Диалог-2003. Протвино, 11–16.06 2003, стр. 373–378.
14. *Kuznetsov I., Kozerenko E.* The system for extracting semantic information from natural language texts// Proceeding of International Conference on Machine Learning. MLMTA-03, Las Vegas US, 23–26 June 2003, p. 75–80.
15. *Кузнецов И. П., Мацкевич А. Г.* Англоязычная версия системы автоматического выявления значимой информации из текстов естественного языка// Труды международной конференции «Диалог 2005», Звенигород, 2005.
16. *Кузнецов И. П., Мацкевич А. Г.* Семантико-ориентированный лингвистический процессор для автоматической формализации автобиографических данных// Труды международной конференции «Диалог 2006», Бекасово, 2006, с. 317–322.
17. *Кузнецов И. П., Мацкевич А. Г.* Семантико-ориентированные системы на основе баз знаний (монография)// М.: МТУСИ, 2007 г.,173с.
18. *Кузнецов И. П. Сомин Н. В.* Англо-русская система извлечения знаний из потоков информации в среде Интернет // Сб. ИПИ РАН, 2007 г.
19. *Кузнецов И. П.* Объектно-ориентированная система, основанная на знаниях в виде XML-представлений// Сб. ИПИ РАН, Вып. 18. 2008 г., с. 96–118
20. *Kuznetsov I. P., Kozerenko E. B.* Linguistic Processor “Semantix” for Knowledge extraction from natural texts in Russia and English// Proceeding of International Conference on Machine Learning, ISAT-2008. 14–18 July,
21. *Segalovich I.* A fast morphological algorithm with unknown word guessing included by a dictionary for a web search engine // MLMTA-2003. http://download.yandex.ru/company/iseg-las-vegas.pdf
22. *А Коваленко.* Вероятностный морфологический анализатор русского и украинского языков// http://www.keva.ru/stemka/stemka.html
23. *Сокирко А. В.* Морфологические модули на сайте www.aot.ru. Диалог-2004. Верхневолжский, 2–7 июня 2004 г. http://www.aot.ru/docs/sokirko/Dialog2004.htm
24. *Сегалович И., Маслов М.* Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов // Диалог'98, Казань, ООО «Хэтер», 1998. http://download.yandex.ru/company/DLG98-MM2.pdf