ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

на тему

Проблемы формализации грамматики и синтаксической неоднозначности при разработке синтаксически размеченного корпуса рассказов Л. Андреева

основная образовательная программа магистратуры по направлению подготовки 45.04.02 «Лингвистика»

Исполнитель:
Обучающийся 2 курса
Образовательной программы
«Прикладная и экспериментальная лингвистика»
очной формы обучения
Ганюкова Мария Игоревна

Научный руководитель:
к.ф.н., ст. преп. Добров А. В.

Рецензент:
лингвист-разработчик ООО «Инфо-Кьюбс»
Попов А. М.

Санкт-Петербург

2018

Оглавление

Введение

В последние годы автоматическая обработка текстов вышла на совершенно новый уровень. Тексты различных стилей и жанров подвергаются автоматическому лингвистическому анализу для самых разных целей. Обработка текстов различных жанров и на естественном языке для удовлетворения широкого спектра потребностей пользователей заняла важное место в области развития технологий.

В частности, синтаксический анализ как один из этапов автоматического анализа текстов играет важную роль в области компьютерной лингвистики. Синтаксический анализ используется в системах машинного перевода, информационного поиска, автоматического реферирования и т.д. Результаты синтаксического анализа художественных текстов могут существенно облегчить работу историков языка и литературоведов, т.к. позволяют отследить и проанализировать функционирование синтаксических конструкций и лексических единиц в них.

Результаты анализа текстов могут быть представлены в виде размеченных корпусов. В настоящее время корпусная лингвистика занимает важное место в области лингвистических исследований. Корпусы сегодня используются для самых разных целей: создание грамматик и словарей, машинное обучение, анализ языкового материала в естественном контексте и т.д. Настоящее исследование посвящено изучению проблем, возникающих при автоматическом синтаксическом анализе текстов при создании корпуса рассказов Л. Андреева.

Актуальность работы определяется повышенным интересом исследователей к вопросам формализации синтаксиса в корпусах текстов. В последние годы ведется активная работа по созданию синтаксически размеченных корпусов (SynTagRus, Hanko, The Penn Treebank Project, The Prague Dependency Treebank и т.д.) и формализации синтаксиса (AOT, RASP, Этап-3 и т.д.). Вместе с тем, в настоящее время исследователи не располагают достаточным количеством

находящихся в свободном доступе систем, позволяющих осуществить качественную и полную синтаксическую разметку корпусов текстов разных стилей, поэтому разработка инструментов, направленных на синтаксический анализ, представляется сейчас актуальной задачей.

Степень разработанности темы может быть охарактеризована следующими положениями.

- Проблемами синтаксического анализа и формализации грамматики занимались многие исследователи (Ю. Д. Апресян, А. В. Гладкий, И. А. Мельчук, Я. Г. Тестелец, Л. Блумфилд, Л. Теньер, Р. Хадсон, Н. Хомский и др.), и в настоящее время существует несколько подходов к представлению структуры предложения и созданию модели синтаксиса. Важной задачей при автоматическом синтаксическом анализе также является снятие омонимии (А. М. Пешковский, О. В. Митренина, М. Шаттлворс и др.).
- Синтаксические конструкции в текстах Л. Андреева в настоящее время изучены либо с позиции рассмотрения моделей с определенной семантикой (И. В. Андреева), либо рассматриваются в качестве одной из ступеней литературоведческого анализа (Е. В. Исаева). Единственной работой в области компьютерной лингвистики, использующей в качестве материала рассказы Л. Андреева, на данный момент является частотный словарь рассказов Л. Андреева (А. О. Гребенников, Г. Я. Мартыненко).
- В настоящее время существуют различные корпусы текстов, в том числе художественных, имеющих синтаксическую разметку и использующих разные подходы к синтаксическому анализу (СинТагРус, Russian Syntax Treebank, ХАНКО и т.д.). Тексты Л. Андреева (в частности, рассказы) представлены только в Национальном корпусе русского языка, но при этом они не вошли в число текстов, имеющих синтаксическую разметку.

Объект исследования – явления синтаксической неоднозначности, возникающей при автоматической разметке, а также явления синтаксиса

художественного текста, которые ранее в недостаточной степени подвергались формализации в существующих формальных грамматиках.

Предметом исследования являются способы формализации грамматики и разрешения неоднозначности художественных текстов, позволяющие повысить полноту и точность автоматической синтаксической разметки корпуса художественных текстов.

Цель исследования: обеспечить повышение точности и полноты в автоматической синтаксической разметке при создании корпуса рассказов Леонида Андреева путем решения проблем, связанных с формализацией грамматики и разрешением синтаксической неоднозначности. Цель исследования предусматривает решение следующих **задач:**

1. Создать коллекцию текстов Л. Андреева, относимых к рассказам, и распределить их по датам написания или первой публикации;
2. Загрузить тексты в корпус-менеджер для автоматической морфологической, синтаксической и семантической разметки;
3. Изучить проблемы, возникающие при синтаксическом анализе текстов Л. Андреева (особое внимание уделяется синтаксису словосочетаний и сверхфразовых единств);
4. Разработать и описать оптимальные способы решения возникающих проблем.
5. Экспериментально апробировать и произвести оценку эффективности разработанных способов повышения полноты и точности синтаксического анализа художественных текстов.

Научная новизна исследования заключается в следующем:

1. Впервые описаны проблемы автоматического синтаксического анализа художественных текстов на материале рассказов Л. Андреева и предложены способы их решения;
2. Создан синтаксически размеченный корпус рассказов Л. Андреева 1900 года.
3. Впервые описаны способы автоматического разрешения

синтаксической неоднозначности на материале рассказов Л. Андреева.

Теоретическая значимость работы определяется тем, что описаны способы формализации синтаксической неоднозначности в ряде конструкций, характерных для художественного текста а) на материале рассказов Л. Андреева; б) с использованием инструментов для автоматической обработки текстов на всех уровнях языка.

Практическая значимость работы заключается в том, что синтаксические конструкции в текстах Л. Андреева свойственны также русскому языку в целом, поэтому способы решения проблем синтаксической неоднозначности и формализации грамматики, разработанные при создании корпуса рассказов Л. Андреева и используемые при синтаксическом и семантическом анализе в проекте AIIRE, могут использоваться и при работе с текстами других авторов и стилей, и применяться в системах автоматического реферирования, информационного поиска, машинного перевода, создания и анализа русскоязычных корпусов, т.к.. Кроме того, результаты работы могут быть полезны при исследовании творчества Леонида Андреева и особенностей художественной литературы Серебряного века.

Материалом для исследования послужили рассказы Л. Андреева 1900 года: «На реке», «Праздник», «Молчание», «Мельком», «Первый гонорар», «Прекрасна жизнь для воскресших», «Рассказ о Сергее Петровиче», «В темную даль», «Ложь».

Методы исследования выбраны с учетом специфики объекта, языкового материала, целей и задач работы. В работе применяются методы классификации, методы лингвистического анализа языкового материала (анализа структур непосредственных составляющих и зависимостей; комплексного анализа синтаксической семантики).

Степень достоверности и апробации результатов: параметром оценки синтаксического анализа является полнота покрытия, т.е. отсутствие нераспознанных единиц, которых нет в морфологическом словаре, и не связанных между собой синтаксических деревьев, — и точность разметки,

проявляющаяся в отсутствии комбинаторных взрывов. Достоверность результатов обусловлена следующими положениями:

- В работе используются как традиционные, так и новейшие отечественные и зарубежные исследования в области компьютерной и корпусной лингвистики;
- Выбранный для анализа материал отвечает целям и задачам исследования.

Результаты исследования были апробированы путем сравнения количества не связанных между собой деревьев и комбинаторных взрывов в начале и в завершении исследования.

Структура работы. Работа состоит из введения, двух глав, заключения, списка использованной литературы, который включает в себя 78 наименований, в том числе 11 на иностранных языках. К работе также прилагаются фрагменты грамматики AIIRE на языке Python, разработанные в ходе исследования (Приложение А). Общий объем работы составляет 92 страницы; основное содержание изложено на 88 страницах, Приложение занимает 4 страницы.

Глава 1. Вопросы автоматического синтаксического анализа в корпусах текстов

1.1 Проблемы лингвистического обеспечения компьютерного синтаксического анализа

Компьютерный синтаксис в настоящее время является одной из самых неоднозначных задач компьютерной лингвистики, вызывающей массу разночтений среди исследователей. В данный момент существуют разные подходы к описанию синтаксиса и различные способы осуществления автоматического синтаксического анализа.

1.1.1 Место компьютерного синтаксиса при автоматическом анализе текстов

Компьютерный синтаксический анализ относится к сфере автоматической обработки текста и занимает важное место в области компьютерной лингвистики. Информация о структуре предложений, о его компонентах и связи между ними требуется для решения многих задач, связанных с машинной обработкой текстов (например, информационного поиска, машинного перевода, автоматического реферирования и т. д.).

Автоматическая обработка текста, помимо синтаксического анализа, включает в себя также морфологический и семантический анализ, преобразуя, таким образом, текст в его «лексемно-морфологическое, синтаксическое и семантическое представление» [Андрющенко 1998: 14]. В широком смысле, однако, автоматическая обработка текста не ограничивается таким представлением и может использоваться в разных областях для различных задач (например, автоматическая обработка текста при автоматизированном редактировании, лексикографической обработке, автоматическом синтезе речи и т.д. – во всех этих задачах преобразование текста имеет свои особенности, направленные на определенные цели). В настоящей работе

термин «автоматическая обработка текста» понимается в более узком смысле: автоматическая обработка текста, используемая при автоматическом лингвистическом анализе текста. При такой обработке синтаксический анализ осуществляется после морфологического, получая на вход его результаты, и перед семантическим, передавая, в свою очередь, результаты синтаксического анализа текста для семантической обработки.

Синтаксический анализ в компьютерной лингвистике называют парсингом – в широком смысле это «автоматический анализ структуры любых текстовых данных» [Добров 2016: 35], в узком – процедура «машинного анализа структуры текста на естественном языке, в том числе – структуры предложения» [Там же: 35]. Результатом работы парсинга в узком смысле является формальное отражение структуры предложения, которое при этом может быть совершенно разным: оно зависит от того, какой подход к представлению грамматики выбирает исследователь. Это может быть дерево зависимостей (предложение как совокупность слов и синтаксические связи между ними) или структура составляющих (предложение как совокупность его частей или словосочетаний). Предложение также может быть представлено в виде структуры, сочетающей в себе элементы как дерева зависимостей, так и структуры составляющих. Такие грамматики называют комбинированными.

В случае представления структуры предложения в виде дерева зависимостей речь идет о грамматиках зависимостей. Такое дерево также называют графом зависимостей. Основоположником грамматики зависимостей принято считать французского исследователя Луи Теньера ([Теньер 1988]). Теньер считал, что «предложение представляет собой организованное целое, элементами которого являются слова» [Теньер 1988: 22], при этом «каждое слово предложения вступает с соседними словами в определенные связи (connexions), совокупность которых составляет костяк, или структуру, предложения» [Там же: 22]. Таким образом, предложение, с точки зрения грамматики зависимостей, состоит из слов и связей между ними;

при этом синтаксическая связь чаще всего оказывается подчинительной, т.е. одно из двух слов является главным, а другое – зависимым. Кроме того, древовидность синтаксической структуры в грамматике зависимостей обуславливается тем, что главное слово является зависимым только по отношению к какому-то одному слову, значит, только одно слово является главным для всего предложения.

При анализе предложения на основе грамматики зависимостей можно столкнуться с некоторыми трудностями.

Во-первых, возникает ряд проблем, касающихся сочинительной связи в предложениях. Сочинительные отношения Теньер рассматривает как особые функции. Суть этих функций сводится к тому, что несколько предложений можно трансформировать в одно, например, *Альфред падает + Бернард падает = Альфред и Бернард падают*. Глагол в последнем случае стоит во множественном числе, т. к. «два первых актанта в единственном числе требуют употребления глагола во множественном числе» [Теньер 1988: 337]. Опровержение идеи использования таких трансформаций как универсального средства для моделирования однородности в русском языке сформировала И.П. Севбо [Севбо 1969], обнаружив, что, например, реципрокальные глаголы не допускают подобных преобразований. Например, при попытке разбить предложение «В душе его боролись желание забыть теперь о несчастном брате и сознание того, что это будет дурно. (Л. Толстой, Анна Каренина)» [Севбо 1969: 17] на два предложения «В душе его боролось желание забыть теперь о несчастном брате» и «В душе его боролось сознание того, что это будет дурно» приводит к потере смысла предложения.

В ряде случаев в грамматику зависимостей вводится специальное сочинительное отношение, но при этом оно устанавливается между самостоятельными словами; союзы при этом не включаются в синтаксическое дерево. При решении ряда задач также предпринимаются попытки разложить сочинительную связь на подчинительные, при этом союз или знак препинания, обозначающий сочинительную связь, можно включить в граф зависимостей

(хотя стоит отметить, что Теньер «старался не включать в синтаксическую структуру предложения служебные слова и знаки препинания» [Теньер 1998: 95]).

Во-вторых, как представляется, наименее убедительное объяснение в грамматике зависимостей получает явление придаточных предложений (в частности, в русском языке особые сложности вызывают придаточные определительные). Например, при построении дерева зависимостей предложения *Алексей Степанович мельком оглядел чердак, другой конец которого утопал в темноте…* [Андреев 1990: 178] лексическая единица *которого* наследует признаки рода и числа от определяемой единицы *чердак*, а падеж – от управляющего им слова конец. Таким образом, одна словоформа зависит сразу от двух, и структура зависимостей при этом не является деревом. Для решения подобных проблем Теньер предлагает такой тип связи как анафорическая связь, с помощью которой выражается тождество между анафорами (словами, отсылающими к элементам предыдущего контекста) и антецедентами (словами, с которыми вступает в связь анафорический элемент, например, предшествующее местоимению имя существительное). Необходимо, однако, учитывать, что «анафора всегда предполагает две семантические связи: 1) связь, которая дублирует структурную связь и 2) дополнительную семантическую связь, которая и составляет анафору» [Теньер 1988: 99]. Тем не менее, необходимо обратить внимание, что в некоторых случаях связь между анафорой и определяемым существительным является также и грамматической (а не только смысловой) (например, в случае, когда определяемое слово является собирательным существительным: «Ленин сказал молодежи, что они должны учиться» [Добров 2016: 39]. В таком случае «… нарушается постулат ГЗ [грамматики зависимостей] о том, что каждое слово может быть грамматически зависимым не более чем от одного слова» [Там же].

В-третьих, при работе с естественным языком нередко возникают случаи, когда слово или словосочетание не может быть привязано к какому-то

одному слову без искажения смысла: для его сохранения слово или словосочетание требует привязки к части предложения или ко всему предложению целиком. Пример такого случая для русского языка приводит А.В. Гладкий: «По графику мы работаем в среду» [Гладкий 1985: 119]. Здесь в зависимости от того, к чему относится по графику, можно понимать предложение по-разному: 1) относится к глаголу работаем: в этом случае можно проинтерпретировать предложение как ‘В среду мы будем работать согласно графику, а в другие дни – нет’; 2) относится к предложению мы работаем в среду: тогда предложение можно понять как ‘График предписывает нам работать в среду, но не в другие дни’.

Таким образом, очевидно, что анализ на основе грамматики зависимостей может вызывать затруднения при применении деревьев зависимостей к конкретному материалу. Я.Г. Тестелец, однако, считает, что «наиболее серьезный недостаток деревьев зависимостей заключается [...] в их неспособности выразить иерархию собственно синтаксических единиц (напомним, что отношение зависимости устанавливается только между словоформами — единицами морфологии)» [Тестелец 2001: 105]. Тестелец утверждает также, что формальная структура составляющих лишена этого недостатка [Там же: 105].

Метод непосредственных составляющих, по определению Ю.Д. Апресяна – это «метод представления словообразовательной структуры слова и синтаксической структуры словосочетания или предложения в виде иерархии вложенных друг в друга элементов» [Апресян 1998: 332]. Он основывается на идее о том, что «всякая сложная единица языка или текста складывается из двух более простых и линейно не пересекающихся единиц – ее непосредственно составляющих (НС)» [Там же: 332]. Так, например, у предложения «Я слышу музыку» [Андреев 1990: 195] можно выделить две НС: 1) именную группу *Я* и 2) глагольную группу *слышу музыку*, состоящую, в свою очередь, из НС слышу и музыку. Следует отметить, что НС *слышу* и *музыку* не входят непосредственно в предложение, но являются компонентами

глагольной группы, входящей в него.

Основные принципы метода непосредственных составляющих были сформулированы Л. Блумфилдом в первой половине XX века. Блумфилд утверждал, что «В любом высказывании языковая форма выступает как составляющее какой-либо более крупной формы, например, John в высказывании John ran away «Джон убежал», либо как независимая форма, не входящая в состав другой, более крупной (комплексной) языковой формы, как, например, John в восклицании John!» [Блумфилд 1968: 178]. В дальнейшем метод получил развитие в трудах Р. Уэллса, З.З. Харриса, Ч.Ф. Хоккета и других американских лингвистов. Позже Н. Хомский в «Логических основах лингвистической теории» [Хомский 1965] внес свой вклад в разработку метода непосредственных составляющих.

В отличие от грамматики зависимостей, грамматика НС объясняет иерархию синтаксических единиц, прибегая при этом к терминам НС (выше уже были упомянуты возможные примеры: именная группа, глагольная группа; однако эти названия могут быть и иными). Помимо этого, грамматика НС объясняет способ порождения предложения: задается ряд правил, позволяющих постепенно «пройтись» по всем НС предложения и в конце концов преобразовать их в линейные цепочки единиц (слов или морфем).

При упомянутых достоинствах, тем не менее, грамматика непосредственных составляющих не лишена недостатков. Самый важный из них заключается в том, что такая грамматика предполагает соответствие между линейным порядком слов и фразовой структурой. Это значит, что при анализе, построенном на НС-грамматике, не учитывается тот факт, что в языках со свободным порядком слов составляющая может разрываться. Например, в предложении - *А как нас знатно вымочило! - продолжал тот же ласковый голос со скрытым смехом* [Андреев 1990: 209] происходит разрыв в составляющей *продолжал со скрытым смехом*. Такая проблема характерна не только для языков со свободным порядком слов, но и для языков со строгим порядком (например, в вопросительных предложениях в английском языке).

Решение было найдено в том, что к правилам переписывания, необходимых для процедуры деривации, добавляются правила перемещения, позволяющие «передвигать» слова в предложении после порождения цепочек с конечными обозначениями.

Стоит, однако, отметить, что описанное решение подходит в большей степени для языков со строгим порядком слов и не находит эффективного применения в русском языке, т.к. трансформации, происходящие с помощью правил перемещения, не работают для языков, подобных русскому. Дж. Р. Росс в своей работе [Ross 1967 (1986)] предполагает, что такое явление как свободный порядок слов возникает в результате поздней трансформации перемешивания, которую Росс предложил считать частью «стилистического компонента» [Ross 1967: 71]. Так, например, в русском языке действуют трансформационные правила, обеспечивающие согласование, падежное оформление и т.д. словоформ. После выполнения этих правил может подключиться «стилистический компонент», который перемешивает изначальную иерархическую структуру предложения. Таким образом, создать работающую формальную грамматику с учетом описанных выше трансформаций не представляется возможным. Стоит также отметить, что в художественных текстах «стилистический компонент» представлен более ярко, чем в новостных сообщениях или научных текстах.

В ходе развития идей грамматик непосредственных составляющих и грамматик зависимостей стало очевидно, что многие проблемы обоих подходов можно решить их комбинированием. Идею о том, что можно создать такую формальную модель предложения, которая сочетала бы в себе преимущества обеих описанных грамматик, развивали в разных подходах такие исследователи как Р. Хадсон [Hudson 1984], В.Б. Борщев, М.В. Хомяков [Борщев, Хомяков 1976]. На материале русского языка комбинированная грамматика была разработана А.В. Гладким. В отношения зависимости в «теории синтаксических групп» Гладкого вступают не слова, а синтаксические группы – «понятие, близкое к понятию составляющей» [Тестелец 2001: 150].

Таким образом Гладкий хотел «получать более естественные – и более близкие к традиционным – описания синтаксической структуры предложений» [Гладкий 1985: 57]. Теория синтаксических групп позволяет при анализе предложения включать в его структуру такие группы, которые вступают в отношения зависимости «целиком», а не посредством одной словоформы. Таким образом, становится возможным, например, анализ предложения *Набросив пальто на голову, он обошел по ледяному, еще не стаявшему пласту вокруг своего домика и заглянул за угол…* [Андреев 1990: 169], где деепричастный оборот *Набросив пальто* на голову присоединяется не к конкретной словоформе, а ко всему предложению в целом.

Комбинированные грамматики на сегодняшний день представляют собой оптимальный вариант при автоматическом синтаксическом анализе текстов, т.к. позволяют описывать структуру предложений наиболее приближенно к тому, как их описывают в традиционных грамматиках, при этом соблюдая строгую формальность. В проекте AIIRE синтаксический анализатор основан на комбинированной грамматике: структуры предложений представлены в виде структур составляющих; при этом в них присутствует информация о зависимостях и линейном порядке единиц.

1.1.2 Неоднозначность и проблема комбинаторного взрыва

Помимо выбора грамматики, оптимально отражающей структуру предложения, перед исследователями в области компьютерного синтаксиса остро стоит проблема неоднозначности языковых единиц. Многозначность слов – лексическая неоднозначность – является самой очевидной, но, помимо нее, может также возникать и морфологическая (многозначность грамматических форм: *море* – именительный падеж слова море или предложный слова *мор*?), и синтаксическая неоднозначность.

Лексическая многозначность, по мнению О.В. Митрениной, являются одной из причин появления синтаксической омонимии. К этой группе можно

отнести такие явления как:

- «омонимия и полисемия»; [Митренина 2005: 6]
- «грамматическая конверсия» [Там же];
- «частичная лексическая омонимия и грамматическая омонимия» [Там же];
- «неоднозначность интерпретации проформы» [Там же].

При этом омография и омофония не рассматриваются в ряде причин, вызывающих лексическую неоднозначность, т.к. «данные явления связаны не с организацией, а с распознаванием синтаксической структуры» [Там же].

В ряде причин, вызывающих синтаксическую неоднозначность, О.В. Митренина также рассматривает «вариативную валентность», к которой относятся «причины, связанные со способностью слова вступать в те или иные синтаксические связи с другими элементами» [Митренина 2005: 7], например: *Доклад об ограблениях в институте социологии* [Там же]; а также группу причин, названную «однородностью»: туда входят «сочетания элементов, провоцирующие возникновение неоднозначных структур, [которые] возникают при наличии однородных членов в одной или нескольких допустимых интерпретациях фразы» [Там же].

Синтаксическая неоднозначность вызывает больше всего проблем при автоматическом синтаксическом анализе. Морфологическая неоднозначность языковых единиц, вариативная валентность и однородность могут приводить к нескольким трактовкам предложения, каждая из которых должна быть представлена в виде отдельной синтаксической структуры. Например, в предложении *В начале поста стрельцы избили его…* [Андреев 1990: 169] существует несколько возможных синтаксических структур, которым соответствуют следующие ситуации:

1. ‘Начало обладает постом’;
2. ‘Начало обозначает процесс’.

Чтобы выбрать из всех вариантов разбора верные, необходимо «...перебрать все комбинации вариантов разбора его [предложения] частей»

[Добров 2016: 44]. При компьютерном анализе это приводит к комбинаторному взрыву. Согласно Криппендорфу, комбинаторный взрыв «происходит, когда посредством увеличения количества объектов, которые могут быть объединены, создается огромное количество возможных комбинаций» [Krippendorff 2010]. Стоит отметить, что комбинаторный взрыв существенно увеличивает время работы машины. Так, например, из-за значительного количества комбинаторных взрывов синтаксический анализ одного абзаца рассказа Л. Андреева «На реке» занял несколько часов, потребовал и использовал существенный объем памяти компьютера. Этот анализ был вынужденно остановлен и так и не был завершен. Для устранения этой проблемы было решено использовать семантический анализ на основе универсальной онтологии проекта AIIRE, что существенно сократило количество версий разбора, объём памяти и количество времени, затраченные на анализ.

Еще одной причиной возникновения комбинаторного взрыва может быть эллипсис – «то есть невыраженность тех фрагментов предложения, значение которых может быть восстановлено из контекста» [Тестелец 2011: 1]. Чтобы обнаружить эллипсис при анализе предложения там, где он действительно есть, необходимо сначала предположить, что он может быть в абсолютно любом месте. Такой подход существенно увеличивает количество комбинаторных взрывов, доводя количество комбинаций до миллионов.

В настоящее время исследователи прибегают к тем способом решения проблемы комбинаторного взрыва, которые соответствуют целям их проектов. В работе [Tomita 1987] описан способ, при котором неоднозначные структуры хранятся в «упакованном» виде [Tomita 1987: 35]. Это значит, что если в предложении есть фрагменты, которые могут быть проанализированы несколькими способами, при этом относятся к одному и тому же классу, но имеют разную структуру, то эти фрагменты в дереве объединяются так, как если бы это был только один узел дерева. Такие узлы называются «упакованными узлами» [Там же]; они позволяют строить одно дерево разбора

для предложений с неоднозначностью, сохраняя при этом возможные варианты разбора.

В работе [Попов, Протопопова, Букия 2016] предлагается выбирать то дерево разбора, в котором разбиения составляющих имеют наибольший вес (что позволяет выбирает наиболее вероятную версию разбора), при этом анализатор «должен строить все варианты синтаксических структур», т.к. «если не перебирать все варианты структур, то нельзя гарантировать, что выдаваемая структура будет иметь наименьший вес». [Попов, Протопопова, Букия 2016: 66]. Упаковка версий разбора в указанной работе является идеологическим продолжением упаковки, описанной в [Tomita 1987], однако имеется ряд отличий. Во-первых, идея М. Томиты предполагает, что для объединения версий в упаковку требуется полное совпадение термов (т.е. должны совпадать в том числе результаты морфологического анализа и токенизации), а в [Попов, Протопопова, Букия 2016] достаточным для объединения условием является совпадение левой и правой границ ветвления составляющей в дереве: например, если есть две версии структурирования одной именной группы с одинаковыми границами, то они объединяются; если границы различаются, то нет. Во-вторых, у М. Томиты при объединении указывается только класс НС, но не грамматические признаки. Это приводит к тому, что могут объединиться составляющие, имеющие разную внешнюю синтактику (например, в именительном падеже NP может быть подлежащим, а в винительном – дополнением), что приводит к некорректному разбору (две падежные версии для подлежащего). Подход, описанный в работе [Попов, Протопопова, Букия 2016], учитывает все признаки, включая грамматические.

Описанные выше решения позволяют экономить память и процессорное время, затраченное на анализ, но при этом проблема неоднозначности остается. Стоит также отметить, что подобный способ избежать комбинаторного взрыва на сегодняшний день работает только на грамматиках с бинарным ветвлением и не может быть использован при работе с грамматиками, допускающими эллипсис и непроективный порядок слов

(такой порядок, при котором предполагается «… дистантное расположение слов, находящихся в отношении непосредственной синтаксической зависимости» [Вороновская 2012:194]). Существует возможность разработать такую грамматику, которая сможет моделировать непроективный порядок слов как проективный и эллипсис как не-эллипсис. Стоит, однако, отметить, что при попытке создать такую грамматику в проекте AIIRE оказалось, что объема памяти недостаточно, чтобы вместить в себя все необходимые правила. Таким образом, этот способ вряд ли подходит в нынешнем виде для применения при анализе художественных текстов.

Уменьшить число версий разбора также можно, прибегнув к эвристическим алгоритмам – это «частные решения, дающие верный или приближенный к верному результат в большей части случаев, но не пригодные для полноценного решения задачи» [Добров 2016: 46]. Такие алгоритмы, однако, редко используются в тех современных системах, где предпочтение отдается многоцелевым парсерам, предлагающим набор версий для каждого предложения. [Там же: 47]. Таким образом, проблема комбинаторного взрыва в настоящее время полностью не решена и требует большого внимания при разработке алгоритмов автоматического синтаксического анализа текстов.

1.1.3 Проблема формализации грамматики

Помимо проблем программного обеспечения компьютерного синтаксического анализа, существует также проблема его лингвистического обеспечения, то есть того, каким образом и насколько полно представлена модель языка в системе автоматического анализа текста.

В проекте AIIRE такая модель, используемая при автоматическом анализе текстов из корпуса, представлена в виде:

1. универсальной онтологии, содержащей информацию о семантике языковых единиц и семантических связях между ними;
2. наборов правил – «грамматики» на языке Python, которые формально

отражают структуру сложных языковых единиц;

3. морфологического словаря;
4. синтаксического модуля, в котором для каждого класса непосредственных составляющих из модуля грамматики прописаны шаблоны семантического графа.

Корпус-менеджер проекта AIIRE «предоставляет возможность просмотра разметки полностью размеченных фрагментов текста» [Гроховский, Добров, Доброва и др. 2017: 158]. Существует три типа помет для фрагментов текста, которые по какой-то причине были размечены не полностью:

1. «Нераспознанные единицы» – «фрагменты, для которых в разметке отсутствуют синтаксические деревья» [Там же: 159];
2. «Разрывы» – «позиции, в которых дерево не может быть связано с соседним» [Там же];
3. «Перекрытия» – «фрагменты текста, в которых пересекаются синтаксические деревья, не полностью покрывающие текст...» [Там же].

Перечисленные инструменты позволяют исследователю вести параллельную работу над разметкой корпуса и совершенствованием формальной модели языка. Пользователь имеет возможность пошагово устранять возникшие проблемы – нераспознанные единицы, разрывы и перекрытия – повышая полноту и точность анализа текста, а также пополняя формальную модель новыми языковыми единицами и явлениями.

Каждая из перечисленных проблем имеет определенное решение.

Проблема нераспознанных единиц возникает, когда между двумя деревьями есть непустой текст, который машина не может распознать и построить для него синтаксическое дерево. В таком случае единица добавляется в морфологический словарь (и тогда либо проблема решена, либо возникает новая – разрыв, перекрытие или комбинаторный взрыв, что влечет за собой дополнительную работу над данным фрагментом текста). К сожалению, добавление одной единицы в словарь не всегда помогает решить

проблему: в морфологии могут отсутствовать две словоформы подряд, или одна словоформа может быть записана так, что получается несколько нераспознанных фрагментов (например, «Дур-р-ра!» в рассказе Л. Андреева «На реке» (1900)). Последняя проблема в настоящее время пока не решена, и подобные единицы (которые часто встречаются в художественных текстах) остаются нераспознанными.

Проблемы разрывов и перекрытий, представляющие особую важность для настоящего исследования, могут быть решены следующими способами:

1. Создание и редактирование выражений и значений в универсальной онтологии, а также моделирование семантических валентностей (установка связей между значениями);
2. Создание и установка новых отношений между значениями в универсальной онтологии, а также расширение и редактирование уже существующих отношений;
3. Пополнение грамматики новыми конструкциями и редактирование уже существующих.

Перекрытия в текстах на русском языке – редкое явление. Таким образом, основной проблемой автоматического синтаксического анализа (по крайней мере, в данном исследовании) стали разрывы и комбинаторные взрывы. Добиться уменьшения их количества — значит приблизиться к созданию полностью синтаксически размеченного корпуса рассказов Леонида Андреева.

1.2 Корпусы с синтаксической разметкой

В настоящем исследовании под лингвистическим, или языковым корпусом текстов понимается «большой, представленный в машиночитаемом виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [Захаров, Богданова 2011: 7].

Сегодня корпусная лингвистика занимает важное место в языкознании. Множество корпусов используется для решения самых разных лингвистических задач: «Ныне создано уже огромное количество корпусов, что определяется многообразием исследовательских и прикладных задач» [Захаров 2016: 141].

Согласно определению В.П. Захарова, «филологическая компетентность» корпуса как массива данных предполагает наличие в нем разметки, которая «заключается в приписывании текстам и их компонентам специальных тегов: собственно лингвистических, описывающих лексические, грамматические и прочие характеристики элементов текста, и внешних, экстралингвистических (сведения об авторе и сведения о тексте: автор, название, год и место издания, жанр, тематика и т.п.)» [Захаров 2016: 143]. По сути, разметка – это то, что делает коллекцию текстов корпусом. Среди типов лингвистической разметки выделяют морфологическую, синтаксическую, семантическую, анафорическую, просодическую и пр.

Это, однако, не значит, что каждый существующий корпус содержит все виды разметки. Ниже будут рассмотрены некоторые корпусы, имеющие синтаксическую разметку.

1.2.1 Зарубежные синтаксические корпусы

Синтаксически аннотированный корпус (Treebank, банк синтаксических конструкций) содержит деревья синтаксического разбора, полученные посредством ручной разметки либо с помощью инструментов синтаксического анализа. Рассмотрим наиболее известные примеры зарубежных синтаксически размеченных корпусов.

The Penn Treebank Project

Банк синтаксических конструкций The Penn Treebank [Marcus 1999] был разработан в Пенсильванском университете и сегодня является наиболее известным примером банка синтаксических деревьев. Разметка этого банка

является в настоящее время стандартом для синтаксического анализа в формализме грамматики непосредственных составляющих. В настоящий момент на сайте представлены синтаксически аннотированные корпусы для разных языков (английского, китайского, корейского, чешского, арабского, а также для исторического корпуса английского языка).

Материалом для корпуса послужили тексты из американской газеты The Wall Street Journal и полностью размеченный Брауновский корпус (первый большой компьютерный корпус). Выдача корпус-менеджера – тексты из этих источников, имеющие морфологическую и синтаксическую разметку.

Морфологическая разметка основана на разметке Брауновского корпуса, однако список тэгов существенно сокращен; к тому же при морфологическом анализе используется синтаксический контекст [Taylor, Marcus 2003: 7]. Результат синтаксического анализа представляет собой иерархическое дерево составных частей высказывания – брэкетинг. В процессе синтаксического анализа в Penn Treebank использовалось два вида брэкетинга: skeletal parsing и predicate-argument structure (предикатно-аргументная структура). В настоящее время при синтаксическом анализе используется второй вид брэкетинга. Подробнее о брэкетинге в Penn Treebank можно прочитать в работе [Taylor, Marcus 2003].

The Prague Dependency Treebank

The Prague Dependency Treebank [Hajic 2006] – один из самых известных синтаксически размеченных корпусов славянских языков, использующий деревья зависимостей. Это корпус текстов на чешском языке, «аннотированный на трех связанных между собой уровнях – морфологическом [...], поверхностно-синтаксическом[...], и глубинно-синтаксическом [...]. На глубинно-синтаксическом уровне аннотируется также актуальное членение предложений и именная кореференция» [Недолужко 2008: 400]. Сценарий разметки корпуса также реализован для словацкого, словенского, греческого, датского и арабского языков.

Корпус включает в себя публицистические тексты из Чешского

национального корпуса объемом 2 миллиона словоупотреблений.

Корпус разрабатывается в Институте формальной и прикладной лингвистики физико-математического факультета Карлова университета в Праге. К корпусу прилагается поисковая программа Netgraph, позволяющая производить поиск по параметрам в корпусе, собирать статистические данные и материал для различных лингвистических исследований.

Разметка поверхностно-синтаксического и глубинно-синтаксического уровня производится вручную на основе предварительной разметки; это значит, что лингвист-эксперт просматривает готовую аннотацию, исправляет и дополняет ее, если это необходимо.

Анализ предложений производится на поверхностно-синтаксическом уровне. Структура предложения представлена «...в виде ориентированного дерева с помеченными связями (ребрами) и узлами. Каждому элементу морфологического уровня соответствует узел поверхностно-синтаксического дерева, отношения между элементами выражены связывающими их ребрами. Тип отношения определяется типом ребра – большинство ребер отражают отношение зависимости, но есть и другие отношения, напр. координация, аппозиция, знаки препинания и др.» [Недолужко 2008: 402]. Каждому узлу приписывается атрибуты, которые в том числе связывают их с элементами морфологического уровня и с глубино-синтаксическим уровнем. Все предложения на этом уровне были сначала проанализированы вручную, а затем на основе предварительной автоматической аннотации. Структура предложения представлена в формализме грамматики зависимостей.

1.2.2 Корпусы русского языка с синтаксической разметкой

Ниже будут рассмотрены наиболее известные корпусы текстов на русском языке, имеющие синтаксическую разметку.

Национальный корпус русского языка (СинТагРус)

Национальный корпус русского языка – это корпус современного

русского языка объемом более 600 млн. словоупотреблений. Часть текстов имеет синтаксическую разметку – эту часть называют синтаксически размеченным корпусом СинТагРус. Он включает в себя общественно-политические, научно-популярные и информационные статьи из журналов и интернет-изданий с 1980 года по настоящее время, а также тексты русской художественной прозы второй половины XX века. Объем корпуса составляет более 1 миллиона словоупотреблений.

Синтаксическая структура предложения в СинТагРус представлена в виде дерева зависимостей, узлами которого являются слова, а ветви помечены именами синтаксических отношений. Разметка осуществляется в полуавтоматическом режиме: сначала происходит морфологический и синтаксический анализ с помощью лингвистического процессора ЭТАП-3, а затем результат этого анализа проверяется и, если необходимо, корректируется лингвистом.

Отличие корпуса СинТагРус от морфологически размеченного фрагмента НКРЯ заключается в том, что «СинТагРус состоит из структур со снятой морфологической и синтаксической омонимией. Это означает, что каждому слову текста сопоставляется единственная морфологическая структура, а каждому предложению ставится в соответствие единственная синтаксическая структура» [Синтаксически размеченный корпус русского языка: информация для пользователей].

Отношения между узлами в дереве зависимостей, отражающем синтаксическую структуру предложения, соответствуют определенным классам синтаксических конструкций. «Особенностью синтаксической модели, на основе которой строятся структуры в данном корпусе, является то, что в ней различается много типов конструкций и, соответственно, используется большое число отношений (67)» [Там же]. Основанием для различения синтаксических отношений являются синтаксические средства (части речи, морфологические характеристики, порядок слов, интонация, знаки препинания, синтаксические и семантические признаки слов).

Russian Syntax Tree Bank

Russian Syntax Tree Bank – это банк синтаксических деревьев, отражающих структуру предложений на русском языке. В него включены результаты разбора 64800 предложений тремя системами автоматического синтаксического анализа: SemSin, SyntAtom и Russian Malt [Russian Syntax Tree Bank]. Материалом послужили научные тексты, современная художественная литература, а также тексты новостных сообщений из Уппсальского корпуса русского языка. Стоит отметить, что сайт RSTB в данный момент недоступен; ссылку на него можно найти на сайте [Синтаксически аннотированные корпуса – NLPub 2012].

Все системы автоматического синтаксического анализа, используемые в корпусе, представляют структуру предложения в виде деревьев зависимостей. Узлами деревьев являются слова предложения. Названия синтаксических связей и морфологические пометы словоформ в корпусе используются те же, что и в исходных системах.

Существует также 800 вручную размеченных предложений этого корпуса – эталонная разметка для сравнения результатов анализа от SemSin, SyntAtom и Russian Malt. Разметка проводилась двумя независимыми аннотаторами в соответствии с инструкцией по ручной разметке.

Корпус на основе Link Grammar Parser

The Link Grammar Parser [Link Grammar] – это синтаксический анализатор английского языка, основанный на грамматике связей – оригинальной теории английского синтаксиса. При анализе предложения система присваивает ему синтаксическую структуру, состоящую из набора маркированных связей, соединяющих пары слов. Анализатор также находит составляющие (именные группы, глагольные группы и т.д.).

Идея этого анализатора была развита для русского языка Сергеем Протасовым, который создал корпусы синтаксически размеченных предложений на русском языке в формализме Link Grammar Parser [Link Grammar for Russian]. Объем первого корпуса составил 30 миллионов

предложений, второго – 21 миллион, третьего – 11 миллионов. Материалом послужили тексты на страницах публичных сайтов в 2007 году.

Грамматика связей – это формализм, который использует связи между последовательностями слов для создания синтаксической структуры. В отличие от грамматики зависимостей, в грамматике связей связи не имеют направления, могут образовывать циклы и разбиваться на иерархические классы, как и сами слова. Корневого слова в грамматике связей нет.

В основе грамматики связей лежит свойство, называемое проективностью и присущее большинству индоевропейских языков: если между словами, которые связаны между собой, провести линии, то эти линии не пересекутся. Грамматика связей состоит из слов, которые имеют ограничения или требования по связям. Последовательность слов является предложением тогда, когда выполнены три следующих правила:

1. Проективность;

2. Связность;

3. Требования (записаны в виде формул в словарях).

Алгоритм анализа предложения «представляет собой рекурсивный разбор предложения сверху вниз с кэшированием промежуточных результатов» [Протасов 2006: 517]. Таким образом, структура предложения представлена в виде связей и графа – соединенных по правилам проективности коннекторов слов. Узлами графа являются слова, дугами – связи с названиями коннекторов. Граф «расположен выше линейно расположенных слов» [Там же].

ХАНКО

ХАНКО – это корпус русских текстов, разработанный в Хельсинском университете в начале 2000-х годов в рамках проекта «Функциональный синтаксис русского языка» (под руководством профессора А. Мустайоки). Материалом для корпуса послужили статьи первых четырех номеров журнала «Итоги» за 2001 год, объем составил около 100 тыс. словоупотреблений. Тексты имею морфологическую, синтаксическую, а также метаразметку

(номер журнала, имя автора, тип текста). Стоит отметить, что в настоящий момент сайт, на котором размещен корпус, недоступен. Ссылку на него можно найти в работе [Копотев, Мустайоки 2003].

Синтаксическая разметка корпуса совмещает разметку в терминах членов предложения (традиционный синтаксис членов предложения) и грамматику зависимостей. Такой подход, по мнению авторов, позволит удовлетворить широкий круг потребностей пользователей и подробно описывать как узлы, так и связи синтаксических структур [Копотев, Гурин 2006].

Особенностью корпуса является «тщательно проработанный формат лингвистического описания данных и полная визуальная (ручная) проверка результатов автоматической разметки, имеющая следствием полное снятие грамматической омонимии, там, где она может быть снята человеком» [Захаров 2015: 25].

1.3 Исследования в области лингвистики на материале текстов Л. Андреева

В качестве материала для настоящего исследования были выбраны рассказы Леонида Андреева. Существует несколько предпосылок для этого решения:

1. Небольшой объем рассказов по сравнению с художественными текстами других жанров позволяет выделить наиболее частотные и характерные для практически всех прозаических произведений писателя синтаксические конструкции, в которых возникают проблемы при автоматическом синтаксическом анализе.
2. Творчество Леонида Андреева в настоящий момент мало изучено (для русской литературы Андреев «воскрес» только во второй половине 70-х годов XX века), так что лингвистические исследования на материале текстов его авторства и создание корпусов этих текстов представляется

перспективным.

Леонид Николаевич Андреев (1871-1919) родился в Орле, где закончил классическую гимназию, а после продолжил учебу в Петербурге, затем в Москве. Получив образование юриста, начал работу в этой области; примерно в то же время стал писать. Первыми произведениями стали репортажи из зала заседаний, которые превращались в сценки. Полученная специальность и работа повлияли на творчество писателя: в его произведениях часто встает вопрос о том, обвинить или защитить человека.

Пик популярности Л. Андреева пришелся на начало 1900-х годов. Многие видели в нем писателя массовой культуры, однако такие писатели быстро забываются. К 1910 году популярность автора пошла на спад.

Он уехал в Финляндию, где после провозглашения ее независимости в 1917 году оказался в изоляции от родной страны. Оказавшись в эмиграции, он в 1919 году умер в полной безвестности.

Особенностью поэтики Л. Андреева является то, что ее сложно охарактеризовать: Андреев и реалист, и модернист, и экзистенциалист, и неореалист и т.д. [Михеичева 1995]. Его творчество эклектично, в нем сочетаются разные методы и направления. С этой точки зрения изучение языка его произведений представляется особенно интересным.

Говорить о творчестве Л. Андреева и о его роли в русской литературе начали только во второй половине 70-х годов XX века. «Воскресение» Андреева связывают с именем Людмилы Иезуитовой, чья книга [Иезуитова 1976] стала одной из первых в современном литературоведении обобщающих работ о творчестве писателя.

1.3.1 Исследования синтаксиса на материале прозы Леонида Андреева

В результате поиска материалов по исследованиям творчества Л. Андреева или с использованием его текстов в качестве материала были обнаружены работы, анализирующие синтаксические конструкции на

материале текстов Л. Андреева: [Андреева 2015], однако оказалось, что в данный момент не существует работ, посвященных автоматическому анализу (в частности, автоматическому синтаксическому анализу) текстов Л. Андреева.

Возрождение Л. Андреева как писателя привело к тому, что в последние 40 лет появились литературоведческие статьи, так или иначе анализирующие творчество и, в частности, рассказы писателя. Особенности авторского синтаксиса, тем не менее, в этих статьях либо не анализируются вообще [Бондарева 2009], [Лукин 2017], либо анализируются крайне поверхностно [Исаева 2011], лишь в той мере, в которой анализ синтаксиса востребован как один из этапов литературоведческого анализа текста. В случаях, когда синтаксису уделяется некоторое внимание, речь идет скорее об общих описаниях, чем об анализе конкретных конструкций и явлений.

Отдельного упоминания заслуживает работа по составлению частотного словаря рассказов Л. Андреева, выполненная в Санкт-Петербургском Государственном Университете А.О. Гребенниковым под редакцией Г.Я. Мартыненко [Гребенников 2003]. Эта работа является первой, в которой тексты Л. Андреева (в частности, рассказы) используются в качестве материала для автоматического анализа – исследования лексико-статистических характеристик текстов.

1.3.2 Представление творчества Леонида Андреева в современных корпусах

В Национальном корпусе русского языка [Национальный корпус русского языка 2003-2018] представлено 62 произведения Л. Андреева различных жанров, из них 46 рассказов. В настоящее время тексты Л. Андреева имеют только морфологическую разметку, и только 4 из них имеют снятую морфологическую омонимию: рассказы «Вор» (1904), «В подвале» (1901), «Кусака» (1901) и «Набат» (1901). Следует отметить, что в ходе

настоящего исследования было собрано 103 рассказа Л. Андреева с 1892 по 1919 года (с опорой на информацию, содержащуюся в шеститомном сборнике произведений Л. Андреева 1990-х годов [Андреев 1990-1995]), однако в корпус-менеджер AIIRE были загружены только рассказы 1900 года.

Следует отметить, что тексты Леонида Андреева практически не представлены в корпусах русского языка (исключением является Национальный корпус русского языка), и тем более в открытом доступе не представлены корпусы текстов этого автора, имеющих синтаксическую разметку. В Машинном фонде русского языка [Машинный фонд русского языка] (проект, направленный на создание большого представительного корпуса русского языка) Л. Андреев также не представлен среди авторов, чьи тексты используются в проекте.

Таким образом, использование текстов Л. Андреева в исследованиях в области компьютерной лингвистики, а также их наличие в корпусах текстов на русском языке представляется недостаточным, что указывает на актуальность настоящей работы.

1.4 Инструменты проекта AIIRE

Проект AIIRE [AIIRE 2015-2018] возник в 2010 году и являлся продолжением проекта OOmink (с 2003 года), изначально нацеленного на создание лингвистического процессора. Проект AIIRE изначально создавался как интеллектуальная информационно-поисковая система, поиск в которой осуществляется путем автоматического понимания текстов и вопросов пользователя. С годами проект развивался, и в настоящее время имеет ряд систем, применяемых для решения различных задач, связанных с обработкой текстов на естественном языке. Среди них можно отметить систему выделения именованных сущностей (AiRecognizer), систему выделения оценочных суждений (AiMiner), систему мониторинга СМИ (СМИРТЕО) и т.д. Основными инструментами, которые помогают осуществить анализ языковых

единиц, являются универсальный лингвистический процессор и универсальная онтология. Тексты, загруженные в корпус-менеджер AIIRE, размечаются при помощи этих инструментов.

Для проведения настоящего исследования были выбраны инструменты и корпус-менеджер AIIRE по следующим причинам:

1. Грамматика AIIRE позволяет работать с непроективным порядком слов и эллипсисом, что особенно важно для работы с художественными текстами;
2. Корпус-менеджер позволяет просматривать фрагменты разметки, которые по какой-то причине не были размечены целиком. Это позволяет устранять проблемы, возникающие при анализе, путем определения причины проблемы и работы с морфологический или синтаксический модулем и универсальной онтологией.

   Ниже рассмотрены основные инструменты анализа проекта AIIRE.

1.4.1 Универсальный лингвистический процессор

Универсальный лингвистический процессор AIIRE – «это библиотека и программная утилита, позволяющая производить полный цикл автоматической обработки текста от байт-последовательности до семантического представления содержания текста» [AIIRE – О проекте 2015-2018]. Работа лингвистического процессора основана на методе межуровневого воздействия, который был предложен Г. С. Цейтиным еще в 80-х годах прошлого века [Цейтин 1985]. Этот метод предполагает анализ текста на всех уровнях одновременно, а не последовательно с передачей каждому уровню результатов анализа предыдущего. Это «позволяет устранять неоднозначность на более низких уровнях, используя правила более высоких уровней еще до того, как неоднозначность более низких уровней успевает привести к так называемому комбинаторному взрыву» [Добров 2015: 115]. Так, в процессе анализа одновременно может быть снята морфологическая

неоднозначность в зависимости от результатов синтаксического связывания или не связывания языковых элементов, а синтаксическая, в свою очередь, – при семантическом несвязывании значений элементов синтаксического дерева.

Таким образом, лингвопроцессор осуществляет одновременный анализ текста на морфологическом, синтаксическом и семантическом уровнях. Ниже подробнее рассмотрен каждый из них.

1.4.1.1 Морфологический анализ

На этапе морфологического анализа происходит определение леммы (т.е. базовой формы слова) словоформы и её грамматических характеристик. Одним из способов морфологического анализа текста является морфологический анализ со словарем словоформ, в который «… можно попытаться заложить [...] все словоформы и для каждой указать ее грамматические значения и другую необходимую информацию» [Бочаров, Митренина 2016: 16]. В проекте AIIRE морфологический анализ выполняется на основе морфологического словаря.

Словарные статьи морфологического словаря AIIRE имеют следующий вид:

```
<entry hw="санкция">
    <inflection template="сущ ru f ina 7a">
        <variable name="основа"
        value="са́нкци"/>
    </inflection>
    <attributes>
        <attr name="lemma" value="санкция"/>
        <attr name="anim" value="0"/>
        <attr name="isName" value="0"/>
        <attr name="pos" value="noun"/>
```

<attr name="gender" value="f"/>

</attributes>

</entry> [Сомс, Добров, Доброва 2014: 153].

Атрибутом элемента *entry* является заголовочное слово словарной статьи (hw – *head word*). Внутри *entry* – два вложенных элемента: *inflection* и *attributes*. Первый содержит информацию о типе склонения слова (при этом используется концепция склонения А.А. Зализняка [Зализняк 2008], а также материалы докторской диссертации С. А. Кузнецова [Кузнецов 2000] в качестве информации о морфологии глагола). Атрибут *template* хранит наименование шаблона словоизменения. Элемент *variable* «позволяет определять значения переменных, используемых при порождении формы слова» [Сомс, Добров, Доброва 2014: 154]. В приведенном выше примере используется основа слова. Если в словоформах происходит чередование, в словарной статье указывается несколько переменных основы для порождения определенных форм слова.

В элементе *attributes* хранится все грамматические (и не только) признаки, общие для всех словоформ данной лексической единицы: лемма, одушевленность, часть речи, род; а также о том, является ли эта единица именем собственным.

Следует отметить, что при работе с аннотированным корпусом текстов в AIIRE существует возможность пополнять морфологический словарь. Так, например, множество словоформ, свойственных художественному тексту и литературному языку Серебряного века (*ныне, буде, чкаю*) не было отражено в словаре. Это стало причиной разрывов при синтаксическом анализе; такие разрывы устранялись путем создания в морфологическом словаре словарных статей для таких единиц.

Морфологический анализ в проекте AIIRE осуществляется двумя способами:

1. Статический. При таком анализе «… по имеющемуся набору единиц (основ, окончаний, корней, приставок, суффиксов и т.д.) заранее

автоматически генерируется вся совокупность словоформ языка, каждой из которых сопоставляются возможные варианты разбора – идентификаторы лексической единицы и совокупности грамматических признаков. Процесс анализа сводится к поиску словоформы в базе данных» [AIIRE – О проекте 2015-2018]. Статистический морфологический анализ используется в настоящем исследовании.

2. Динамический. При таком анализе «цепочки алфавитных символов сегментируются всеми допустимыми способами, таким образом, чтобы каждый сегмент соответствовал известной системе морфологической единице (основе, окончанию, приставке и т.д.) В процессе сегментации выполняется связывание выделенных единиц, с алгоритмической точки зрения эквивалентное синтаксическому связыванию. Результатом анализа является совокупность гипотез связывания всех обнаруженных атомарных единиц в единые комплексы, снабжаемые грамматической информацией» [Там же]. Стоит отметить, что динамический морфологический анализ в настоящее время не применяется для русского языка. Он реализован для английского, арабского и тибетского языков.

Морфологический анализ в AIIRE происходит одновременно с другими видами анализа (синтаксическим и семантическим), что частично позволяет избежать проблемы комбинаторного взрыва. Языковые единицы, полученные в результате морфологического анализа, подвергаются синтаксическому связыванию со всеми непосредственными левыми соседями (словоформами, словосочетаниями, предложениями).

1.4.1.2 Синтаксический анализ

В системе AIIRE синтаксический анализ – «это процедура, позволяющая производить переход от синтаксически неделимых единиц к синтаксическим структурам всей анализируемой цепочки» [AIIRE – О проекте

2015-2018]. При работе с корпусом рассказов Л. Андреева неделимыми единицами синтаксического анализа выступали словоформы; при динамическом морфологическом анализе таковыми являются корни и аффиксы.

С точки зрения представления грамматики, в системе AIIRE используется так называемая комбинированная грамматика. Синтаксические структуры представлены в виде размеченных структур составляющих, то есть частей предложения, которым приписана информация о их зависимостях и линейном порядке в предложении. В модуле синтаксической семантики для каждого класса непосредственных составляющих (описывающего синтаксическую конструкцию и ориентированную на «… построение нестрого бинарных комбинированных структур составляющих и зависимостей, допускающих разрывную линеаризацию» [Добров 2014: 173]) прописан шаблон семантического графа (т.е. информация о том, как построить семантический граф из значений главной и зависимой составляющих).

Синтаксический модуль (он же – грамматика AIIRE) представлен в виде иерархически выстроенных классов, написанных на языке Python. Например, класс, описывающий сочетания именной группы с обособленным приложением единственного числа, выглядит следующим образом:

```
class InstanceWithDetachedAdjunct (SyNode):
    head_classes = ['Instance']
    modifier_classes = ['DetachedAdjunct']
    modifier_can_be_ellipsed = 0
    head_can_be_ellipsed = 0
    specifier_order = Orders.Right()
    omit_grammemes = ['isName', 'nameRole']
```

Следует отметить, что не все возможные признаки отражены в примере; и не все признаки, указанные в примере, могут быть указаны для других классов. Рассмотрим основные признаки, используемые для описания классов в грамматике.

SyNode является общим надклассом для всех классов НС и может использоваться только для тех составляющих, которые являются главными по отношению к другим (зависящим от них) составляющим. Признак *head_classes* указывает на возможные классы главной дочерней составляющей. В примере выше таким классом является класс Instance – класс для характеристики референтной именной группы, обозначающей конкретный предмет или явление в мире. *modifier_classes* указан для тех классов непосредственных составляющих, которые могут быть модификаторами непосредственно по отношению к указанному классу. Признак *modifier_can_be_ellipsed* указывает на то, может ли НС, указанная как модификатор, быть эллиптирована; в приведенном примере такого не может быть, т.к. если эллиптируется приложение, то невозможна становится конструкция именной группы, сочетающейся с приложением. Признак *head_can_be_ellipsed* означает, что ядро в описываемой НС может отсутствовать; в данном случае это невозможно, т.к. без ядра получится приложение в запятых, которое не может брать на себя функцию именной группы. Признак *specifier_order* указывает на порядок зависимой НС относительно главной; в данном случае приложение должно находится справа от именной группы; также в грамматике прописаны составляющие, находящиеся от главной НС слева (Orders_Left()) и как справа, так и слева (Orders_Optional). Признак *omit_grammemes* содержит массив грамматических признаков, которые не передаются от главного слова к родителям и от родителя к модификаторам.

Таким образом, за счет выстраивания иерархии для всех составляющих в грамматике, указания их признаков и информации о зависимостях, становится возможным при синтаксическом анализе производить процедуру синтаксического связывания единиц. «Эта процедура состоит в поиске всех возможных маршрутов в грамматике составляющих, связывающих две единицы и имеющих одну точку перелома. В случае отсутствия таких маршрутов связывания не происходит и единицы считаются несвязанными»

[AIIRE – О проекте 2015-2018]. Если единицы успешно связались между собой, они подвергаются семантическому анализу; в случае успешного прохождения семантического анализа структура подвергается связыванию с соседними единицами.

В случае несвязывния единиц перед исследователем стоит задача определить, на каком уровне (морфологическом, синтаксическом, семантическом) возникла проблема, и устранить, таким образом, морфологическую, синтаксическую или семантическую неоднозначность.

1.4.1.3 Семантический анализ и универсальная онтология

Семантический анализ в проекте AIIRE – «это процедура, направленная на вычисление семантики входного потока на основе результатов его синтаксического анализа» [AIIRE – О проекте 2015-2018]. Входной единицей для семантического анализа принято считать атомарную единицу или конструкцию, состоящую из нескольких единиц. Семантический анализ атомарных единиц происходит путем извлечения понятий, соответствующих этим единицам, из Базы Знаний (универсальной онтологии) AIIRE. Анализ конструкций производится путем вычисления семантического графа (при этом используются понятия, которые были получены при семантическом анализе дочерних узлов конструкции) в соответствии с правилами, установленными в грамматике для каждого класса. При этом, если построенный граф противоречит правилам об отношениях, установленным в грамматике, он считается неверным. Таким образом происходит снятие различных видов неоднозначности на всех уровнях анализа.

Система AIIRE предполагает проведение автоматического анализа текста в том числе и без семантического анализа. Учет семантики единиц и отношений между ними, однако, существенно ускоряет процесс анализа и уменьшает количество комбинаторных взрывов. Вместе с тем, при работе с корпусом текстов Л. Андреева оказалось, что существенное количество

единиц не имеют отражения в Базе Знаний, что стало одной из проблем возникновения разрывов при синтаксическом анализе.

База данных проекта AIIRE, содержащая значения слов, называется онтологией, хотя также ее можно было бы назвать и семантическим словарем, и тезаурусом, и базой знаний. В [Dobrov 2014] приводится несколько причин, почему используется именно термин онтология:

1. Семантические словари зависят от языка, тогда как онтологии – нет. Онтология AIIRE, помимо лексических значений, содержит такие понятия, которые не могут быть привязаны к конкретным лексическим единицам (например, единое понятие ‘объект или процесс’) [Dobrov 2014: 148];
2. Тезаурусы, такие как, например, WordNet [Fellbaum 1988], являются семантическими словарями, однако они ограничены взаимосогласованными отношениями (синонимия, гиперонимия и пр.). Тезаурусы не содержат концептуальных отношений, основанных на экстралингвистических знаниях, тогда как в онтологии AIIRE они представлены. Например, отношение ‘может выполнять действие’ может помочь снять лексическую неоднозначность в предложении The table was moved to the corner of the room, т.к. в отношении отражена информация о том, что только физические объекты могут перемещаться в физическом пространстве [Dobrov 2014: 149];
3. База знаний – это, вероятно, более подходящий термин для обозначения онтологии AIIRE, чем словарь или тезаурус, но все же он недостаточно специфичен, чтобы отражать тот факт, что элементы, хранящиеся в онтологии, являются моделями понятий, которые образуют не только иерархию наследования, но также и вышеупомянутые отношения. Онтологии иногда даже рассматриваются как виды баз знаний [Knowledge Base 2014], которые, в отличие от других видов баз знаний, имеют

иерархическую структуру [Dobrov 2014: 149].

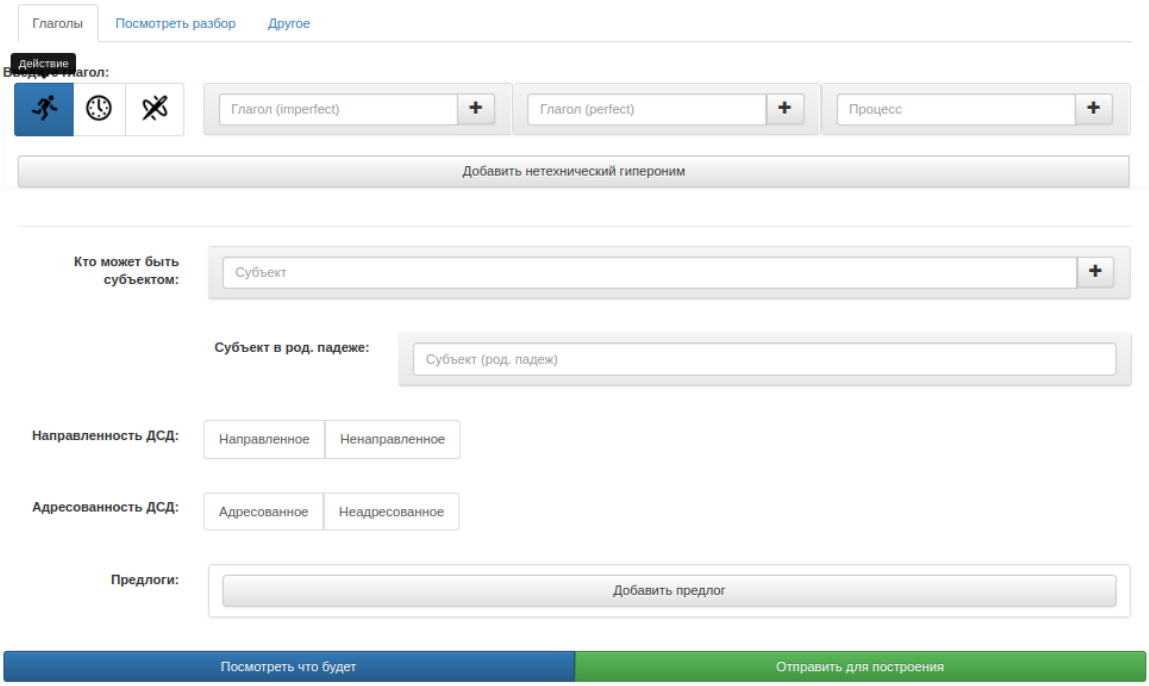
Необходимо также отметить, что одной из особенностей онтологии AIIRE является то, что отношения, устанавливаемые между концептами, сами являются концептами и образуют собственную иерархию наследования. Отношения, как и любые концепты, могут также вступать в другие отношения.

Концептами в AIIRE называются значения слов, представленные в виде набора атрибутов, каждый из которых представляет пару «отношение – объект». В каждом концепте указано лексическое значение языковой единицы, ее гиперонимы, гипомины и синонимы, а также объекты и субъекты отношения (для концептов-отношений, например, ‘обладать совокупностью’).

В используемой онтологии «существуют строгие правила наследования и замещения отношений. Если один концепт наследует другой, то каждый атрибут наследуемого воспроизводится наследующим концептом, однако наследуемый атрибут может быть замещен другим атрибутом, если и отношение, и объект замещаемого атрибута наследуются или совпадают с отношением и объектом замещающего» [Добров 2015: 115]. Для экономии вычислительных ресурсов в онтологии прописаны только прямые отношения, обратные же вычисляются.

Отдельно стоит сказать об обработке глаголов в онтологии AIIRE. Ю.С. Маслов в Лингвистическом энциклопедическом словаре утверждает, что глаголы делятся на динамические (предельные — «обозначают действия, направленные к пределу и исчерпывающие себя с его достижением» и непредельные — «обозначают действия, не предусматривающие предела в своем протекании») и статические [Маслов 1998: 105]. Статические глаголы обозначают «состояния, зависящие от воли субъекта (стою) либо не зависящие от нее (болею, мерзну), отношения (соответствует, превосходит), проявления качеств и свойств (трава зеленеет…)» [Там же: 105]. Такое различие легло в основу создания онтологичекого редактора ONTOHELPER [ONTOHELPER], с помощью которого в онтологии создаются концепты глаголов (рис. 1)

Рисунок 1 Онторедактор ONTOHELPER для добавления в базу знаний глагольных значений



При использовании редактора ONTOHELPER необходимо определить, обозначает обрабатываемый глагол действие (предельные глаголы), состояние (статические) или деятельность (непредельные). Для предельных глаголов указывают их видовые пары и существительное, обозначающее процесс по этому действию. Существительное также указывается для непредельных и статических глаголов. Таким образом, в онтологии создается три или два концепта соответственно. Каждый из них сохраняет указанные валентности (то есть способность сочетаться с другими лексическими единицами), в том числе с предлогами. Указывается направленность (переходность) глагола, адресованность (управление дативом), а также классы предлогов. Все эти поля в онторедакторе обязательны для заполнения.

Результат семантического анализа представляется в виде семантического графа, ребра которого «могут быть вершинами и иметь собственные ребра, при этом вершины и ребра СГ [семантического графа] — концепты онтологии или наследующие их текстовые концепты» [Добров 2015: 166].

1.5 Выводы

1. Материалом для исследования были выбраны рассказы Л. Андреева, т.к., во-первых, тексты его произведений представлены только в одном корпусе русских текстов (Национальный корпус русского языка) в неполном объеме; во-вторых, в настоящее время нет исследований, посвященных автоматическому синтаксическому анализу художественных текстов на материале произведений этого автора; в-третьих, исследовать творчество писателя стали относительно недавно (70-е года прошлого века), поэтому результаты настоящего исследования представляются перспективными для использования при изучении текстов Л. Андреева;
2. Инструментом для настоящего исследования послужил проект AIIRE, находящийся в свободном доступе и позволяющий создавать корпусы текстов и проводить автоматическую морфологическую, синтаксическую и семантическую разметку; а также просматривать элементы текста, в которых разбор по какой-либо причине не произошел, и устранять эти проблемы в ходе исследования. Грамматика проекта составлена таким образом, что позволяет работать с непроективным порядком слов в предложении и учитывать эллипсис, что представляется особенно важным при анализе художественных текстов;
3. Представление структуры предложения в проекте AIIRE и, соответственно, при синтаксическом анализе рассказов Л. Андреева происходит в формализме комбинированных грамматик: в модуле синтаксической семантики прописан шаблон семантического графа для каждой составляющей;
4. При синтаксическом анализе было решено использовать также семантическую разметку, т.к. это существенно сокращает количество

разборов и за счет этого уменьшает число комбинаторных взрывов (существенно сокращая, тем самым, время, затраченное на разбор, и объем занимаемой памяти). Без использования семантической разметки разбор одного абзаца текста занимает несколько часов.

Глава 2. Синтаксические конструкции, вызывающие проблемы при автоматическом синтаксическом анализе, и способы решения этих проблем

Корпус-менеджер проекта AIIRE предоставляет исследователю возможность просмотреть фрагменты разметки текста. Для фрагментов, которые были размечены частично, существуют пометы: разрывы, нераспознанные единицы и перекрытия (см. параграф 1.1.3 настоящей работы). Нераспознанные единицы, разрывы, а также комбинаторные взрывы стали проблемами на пути к полной и точной синтаксической разметке рассказов Л. Андреева.

Для устранения той или иной проблемы, проявляющейся при синтаксическом анализе, необходимо проанализировать синтаксические конструкции, в которых они возникают, а также единицы, из которых состоят эти конструкции (так, например, некоторые лексические единицы могут отсутствовать в морфологическом словаре и значения некоторых из них могут не иметь представления в универсальной онтологии).

При анализе частично размеченного текста вручную были отобраны синтаксические конструкции, проблемы при анализе которых характеризовались следующими положениями:

1. Проблема при анализе конструкции носит систематический характер, т.е. проявляется в одном и том же месте в каждой повторяющейся конструкции;
2. Синтаксическая конструкция, при анализе которой возникает проблема, характерна для художественного текста.

Таким образом удалось отобрать несколько синтаксических конструкций, характерных для художественного текста, ошибки при синтаксическом анализе в которых возникают систематически. Стоит, однако, отметить, что в рамках настоящего исследования удалось решить проблемы не со всеми отобранными конструкциями – предполагается, что работа с ними

будет продолжаться при дальнейшем исследовании проблем автоматического синтаксического анализа на материале текстов Л. Андреева.

2.1 Корпус рассказов Леонида Андреева

Для создания корпуса рассказов Л. Андреева использовался корпус-менеджер проекта AIIRE, позволяющий при помощи лингвистического процессора разметить тексты (морфологическая, синтаксическая разметка и возможность подключить семантическую разметку), а также просматривать и анализировать полученные результаты.

На первом этапе исследования – сборе текстов – возник ряд проблем:

1. Не существует четких критериев для определения жанра произведения (например, такие произведения как «Рассказ о семи повешенных» (1908) или «Жизнь Василия Фивейского» (1903) в разных источниках могут относится или к рассказам, или к повестям);
2. Разные источники предлагают различное распределение текстов по годам (по году написания или году первой публикации, при этом год написания также может различаться);
3. Не все тексты рассказов представлены в электронном виде.

Описанные проблемы были решены следующим образом: тексты, относимые к жанру рассказов, и распределение их по годам отбирались с опорой на собрание сочинений Л. Андреева в 6-ти томах 1990-х годов (тома 1, 2, 3, 4, 5) [Андреев 1990-1995], а также на полное собрание сочинений Л. Андреева в 23-х томах (1, 5 и 6 тома), выпуск которого в данный момент еще не завершен [Андреев 2007-2013]. Из этих же источников были взяты тексты рассказов, не представленные в машиночитаемом виде (сами собрания сочинений были представлены в формате DjVu) и обработаны в программе FineReader от компании ABBYY [Сервис распознавания текста онлайн 2008-2018]. Данная программа позволяет конвертировать отсканированные файлы и файлы в формате PDF в редактируемые форматы. Тем не менее, не все

фрагменты текста были распознаны, поэтому тексты, прошедшие обработку в FineReader, были затем вручную отредактированы.

При загрузке текстов в корпус-менеджер AIIRE обнаружилось, что при анализе возникает существенное количество разрывов, т.к. ранее в проекте не производился автоматический анализ художественных текстов. Обнаружилась необходимость добавления в грамматику правил для новых составляющих, а в морфологический словарь – новых единиц. Таким образом, было принято решение в рамках настоящего исследования ограничить количество загружаемых текстов и решить при работе с ними наиболее распространенные проблемы автоматического синтаксического анализа. Для работы были выбраны рассказы 1900 года по следующим причинам:

1. Начало 1900-х годов характеризуется пиком популярности Л. Андреева;
2. Количество рассказов, написанных в 1900 году (9) представляется достаточным для выявления наиболее частотных проблемных конструкций, характерных для текстов Л. Андреева по крайне мере начала 1900-х годов. При дальнейшей разработке темы возможна сравнительная характеристика особенностей синтаксиса текстов Андреева на материале рассказов разных годов.

Таким образом, в корпус вошли следующие рассказы: «На реке», «Праздник», «Молчание», «Мельком», «Первый гонорар», «Прекрасна жизнь для воскресших», «Рассказ о Сергее Петровиче», «В темную даль», «Ложь», написанные в 1900 году. Объем корпуса составил 36093 словоупотреблений. Корпус доступен по ссылке [Корпус рассказов Л. Андреева 2017-2018].

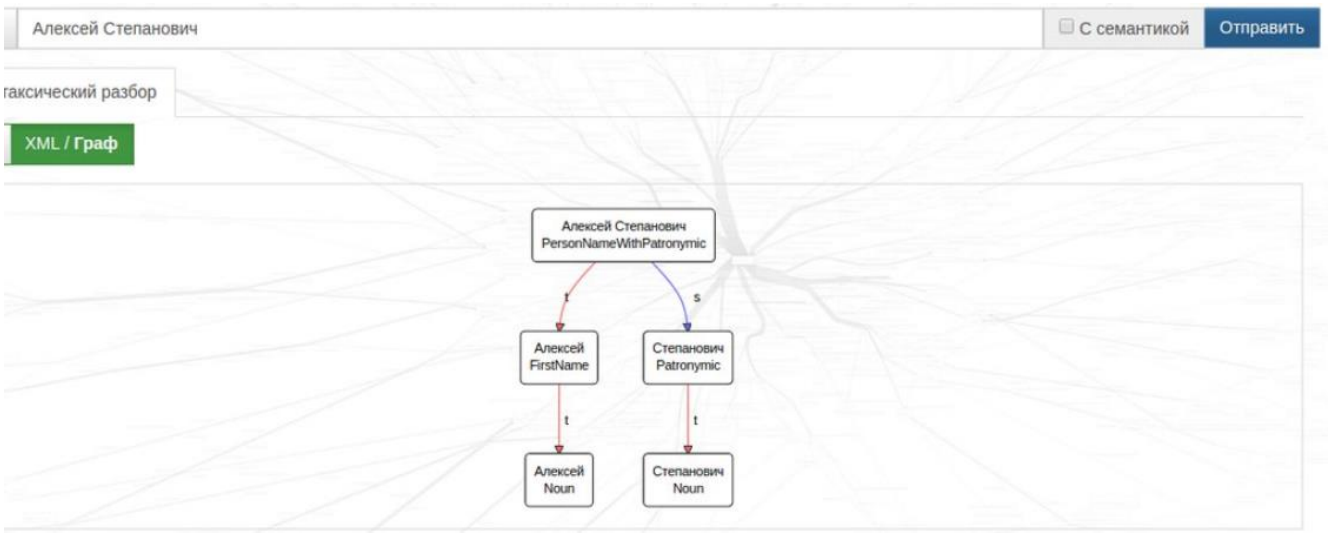
2.2 Обособленные приложения

Приложение – это «...определение, выраженное именем существительным. Термин «П.» [приложение] указывает также на особый характер связи между определяемым и определяющим – их согласование на началах параллелизма» [Кручинина 1998: 398]. Приложения по отношению к

определяемому слову могут обозначать «...качества и свойства [...], эмоциональные характеристики и оценки [...], функциональные признаки [...], видовые признаки [...], имена, фамилии, клички, прозвища [...], географич. названия [...], условные наименования (собств. имена) предметов [...]» [Там же]. В настоящей работе рассматриваются приложения, отделяемые от определяемого слова запятой, как в следующих примерах: *Алексей Степанович, машинист при Буковской мельнице, среди ночи проснулся…*; *Сухо поздоровавшись с управляющим, толстым и рыжим мужчиною, Алексей Степанович стал смотреть на реку.* Такие приложения называют обособленными.

При работе с текстами рассказов Л. Андреева в корпус-менеджере AIIRE было замечено, что возникают разрывы между именными группами и группами обособленных приложений, которые зависят от них. Так, например, в предложении *Алексей Степанович, машинист при Буковской мельнице, среди ночи проснулся...* возникает разрыв между лексический единицей *Степанович* и запятой. Было выявлено, что этот разрыв связан с отсутствием в грамматике AIIRE правил, описывающих приложение: словосочетание *Алексей Степанович, машинист при Буковской мельнице,...* было введено в лингвопроцессор с выключенным модулем семантического анализа, при этом проверялся каждый разрыв по очереди. Так, например, несмотря на наличие разрыва, было построено дерево для словосочетания *Алексей Степанович* (см. рис. 2).

Рисунок 2 Синтаксическое дерево разбора словосочетания *Алексей Степанович*



Построение в режиме выключенного семантического анализа синтаксического дерева для элементов, между которыми в противном случае возникает разрыв, означает, что причина разрыва кроется в семантике – вероятно, одно из этих слов было некорректно обработано в онтологии. Для словосочетания *Алексей Степанович, машинист* не было построено единого связного дерева, что означало, что разрыв между определяемой составляющей *Алексей Степанович* и обособленным от нее приложением обусловлен отсутствием необходимой конструкции в грамматике.

Для создания конструкции, описывающей функционирование обособленного приложения в предложении, необходимо было прописать классы для отсутствующих в грамматике составляющих на языке Python.

1. Класс InstanceWithDetachedAdjunct описывает обособленное приложение единственного числа, которое согласуется с определяемым существительным в роде, числе и падеже, например: *Алексей Степанович, машинист при Буковской мельнице…*;
2. Класс DetachedAdjunct, описывающий приложение в единственном числе, обособленное запятыми, например*: , машинист при Буковской мельнице,*;
3. Класс OpenDetachedAdjunct для описания фрагмента обособленного приложения, не имеющего «закрывающей» запятой: *, машинист при*

Буковской мельнице;

4. Класс Adjunct, описывающий необособленное приложение. Например, в словосочетании *Алексей Степанович, машинист при Буковской мельнице*, классу Adjunct будет соответствовать НС *машинист при Буковской мельнице*.

При этом классы InstanceWithDetachedAdjunct и OpenDetachedAdjunct являются подклассами класса SyNode и являются главными по отношению к классам DetachedAdjunct и Adjunct соответственно, а класс DetachedAdjunct – модификатором по отношению к классу InstanceWithDetachedAdjunct (под модификатором «...подразумеваются элементы значения предложения, соответствующие значениям зависимых узлов и не являющиеся аргументами главных» [Добров 2014: 173]. Здесь важно отметить ряд особенностей наследования признаков в грамматике AIIRE:

1. У каждой непосредственной составляющей есть свои грамматические признаки. Признаки родительской составляющей копируются от ядра – «… дочерней составляющей, помеченной как главная» [Там же] (это значит, например, что класс InstanceWithDetachedAdjunct или Adjunct наследуют признаки класса Instance, (класс для характеристики референтной именной группы, обозначающей конкретный предмет или явление в мире), указанного для них в массиве *head_classes*);
2. Признаки родителя передаются модификатору (ядро и модификатор согласуются через родителя). Так, например, класс DetachedAdjunct наследует признаки класса InstanceWithDetachedAdjunct и ядра Instance;
3. Признаки не наследуются, если они указаны в массиве *omit_grammemes*;
4. Признаки не передаются аргументу (этим аргумент отличается от модификатора), если они не указаны в массиве *delegates_grammems*. В противном случае они не передаются аргументу, но не передатся ядру;
5. Родитель может требовать определенных признаков от ядра (указываются в словаре *needs_gram*), но за счет согласования с модификатором он также может косвенно требовать их через его ядро.

Таким образом, все введенные классы непосредственных составляющих наследуют признаки именной группы (anim (одушевленность), case (падеж), cop (признак прилагательного, определяющий возможность употребляться только при связочном глаголе (в краткой форме) или не только при нем (в полной форме)), gender (род), num (число), pers (лицо)), кроме признаков isName и nameRole, характеризующих слово с точки зрения его принадлежности к именам собственным и роли имени собственного в соответствующей конструкции, обозначающей именованную сущность (первое имя, отчество, фамилия, название организации и т.д.) соответственно. Эти признаки не наследуются классом InstanceWithDetachedAdjunct (указаны для него в массиве omit_grammems), т.к. обособленное приложение в согласовании с существительным не может выступать в роли наименования в конструкции, обозначающей именованную сущность.

Для всех введенных составляющих также был указан линейный порядок в предложении зависимой НС относительно главной, а также необходимые знаки препинания в качестве аргументов (в массиве *argument_classes*) для классов DetachedAdjunct и OpenDetachedAdjunct.

Классы для непосредственных составляющих, введенные в грамматику AIIRE для решения проблемы разрыва между существительным и зависящим от него обособленным приложением, представлены в Приложении А.

2.3 Сочетания именных групп

Родительный падеж занимает в русском языке одно из первых мест по частоте встречаемости у именных групп [Слюсарь, Самойлова 2015]. Особенно конструкции с родительным падежом характерны для художественных текстов: А. Дубовик в работе [Дубовик 2017] приводит следующие характеристики текстов художественного стиля:

- «лексическое богатство» [Дубовик 2017: 30];
- «употребление преимущественно семантически конкретных

существительных» [Там же];

- «высокая частотность форм именительного и родительного падежей имен существительных» [Там же];
- «преимущественное использование простых словосочетаний» [Там же: 31].

При отборе наиболее частых конструкций, в которых систематически возникали разрывы, самой большой группой оказались сочетания одной именной группы с другой, стоящей в родительном падеже. При устранении разрывов в таких конструкциях они были разделены на несколько подгрупп в зависимости от того, что вызвало разрыв, и способа его устранения.

2.3.1 Сочетания именных групп со значением процессуальности

Первый класс случаев, выделенный из множества именных групп, сочетающихся с именной группой в родительном падеже, состоит из словосочетаний, в которых главной составляющей является группа имени существительного со значением действия, деятельности или состояния, по значению соотносимое с глаголом, например: *шум дождя; звуки голосов; чувство тошноты*. «Грамматика русского языка» предлагает два вида отношений в таких словосочетаниях:

1. Определительно-субъектные отношения: «... отношения между действием, называемым главным словом, и производителем действия, называемым зависимым словом… [...] Зависимое слово в таких словосочетаниях может называть как лицо или конкретный предмет, так и отвлеченное понятие» [Виноградов, Истрина 1960: 238].
2. Определительно-объектные отношения: в них «... главное слово может быть выражено именем существительным, соотносительным с переходным глаголом, требующим винительного падежа без предлога» [Там же]. К этому же виду относятся «словосочетания с главным словом - именем существительным, соотносительным с глаголом, требующим

родительного падежа…» [Там же], например: желание деятельности.

Ниже описан процесс решения проблемы разрывов для каждого класса.

2.3.1.1 Словосочетания с определительно-объектными отношениями

Словосочетания, в которых главное слово выражено существительным, которое соотносится с переходным глаголом или глаголом, требующим родительного падежа, встретились в рассказах Л. Андреева, например, в следующих контекстах:

- «Но было одинаково тяжело и то и другое и вызывало одинаковое *чувство тошноты и скуки*» [Андреев 1990: 170] (курсив наш. - М. Г.).
- «Звонкие голоса мельников, ловивших доски, сияние неба и солнца, разноголосый крик на той стороне, казавшийся так же веселым под этим чистым небом, белый клубочек дыма - все это создавало живую и радостную картину и наполняло душу бодростью и *желанием деятельности*, такой же живой и веселой» [Там же: 172] (курсив наш. - М. Г.).

В первую очередь при решении проблемы с разрывами в словосочетаниях с определительно-объектными отношениями были добавлены отсутствующие в онтологии необходимые концепты и исправлены уже существующие в ней, но добавленные автоматически или неверно обработанные. Следует отметить, что эта стадия работы над устранением разрывов имеет место при работе с каждой проблемной группой, т. к. многие слова, употребляющиеся в художественном тексте, не встречаются в текстах новостных сообщений, на анализ которых изначально были настроены инструменты AIIRE. Именно поэтому пополнение онтологии является неотъемлемым пунктом в описываемой работе.

Концепты для значений лексических единиц *тошнота* и *скука* не были утверждены лингвистами в онтологии (это значит, что концепты не были проверены лингвистами и не допущены к использованию при семантическом

анализе), следовательно, они отсутствовали в той версии онтологии, которая используется лингвистическим процессором. Вместо них были построены автоматические концепты, но только для использования лингвистическим процессором, в самой же онтологии автоматических концептов такого рода нет. Следовательно, в лингвистическом процессоре эти концепты не имели должного описания значений. С помощью онторедактора для глаголов ONTOHELPER были обработаны глаголы *тошнить* и *скучать* (поскольку приведенные существительные соотносятся по значениям с этими глаголами). При этом при обработке глагола *тошнить*, значение которого было определено как состояние, стало ясно, что для формального заполнения субъектной валентности классов значений подобных ему безличных глаголов потребовалось создание фиктивного концепта (для субъектов состояния). Таким образом, в онтологию был добавлен концепт ‘фиктивный субъект для безличного глагола’, который является объектом отношения ‘участие в роли субъекта действия’ ‘осуществлять/осуществить действие или состояние фиктивного субъекта для безличного глагола’, благодаря чему стало возможным обработать в онторедакторе безличный глагол, для которого в онторедаторе в обязательном порядке требуется указание на субъект действия, деятельности или состояния.

При обработке глаголов *тошнить* и *скучать* в ONTOHELPER были указаны существительные *тошнота* и *скука* в качестве процессуальных существительных, соответствующих этим глаголам. При этом для обоих состояний было указано, что они являются неадресованными, состояние тошнить направлено на живое существо (*кого-то тошнит; меня тошнит*); глагол же скучать - ненаправленный.

Следующим шагом стало редактирование выражений *чувствовать* и *желать* в ONTOHELPER, т.к. их изначальная обработка представлялась не вполне корректной. Так, глагол *чувствовать* был разобран как действие. По-видимому, к этому привело добавление глагола *почувствовать* как указание на завершение этого действия. Данная модель значения этого глагола в

онтологии, как представляется, не являлась корректной, поэтому значение глагола *чувствовать* было заново обработано как состояние, для описания процесса которого используется существительное *чувствование* (в онтологии значение этого существительного является строгим синонимом значения существительного *чувство*, т.е. должны совпадать в том числе их валентности). Таким образом, после редактирования глагол *чувствовать* со значением ‘ощущать некоторые чувства’ обладает субъектом ‘живое существо’, является неадресованным и направленным на ‘объект или процесс’.

Глагол *желать* также был представлен в онтологии, однако был обработан как модальный, поэтому для разбора были доступны только его сочетания с инфинитивом: *желать прийти, желать поесть, желать поспать*. Таким образом, при разборе словосочетания *желание деятельности* возникал разрыв. Для его устранения необходимо было внести в данное выражение такое значение, которое делало бы возможным разбор глагола *желать* в сочетании с существительными, т.к. в проекте AIIRE принята конвенция считать разными значения пусть похожие, однако различающиеся по валентностям. Таким образом, были созданы два значения для этого глагола:

1. ‘желать чего-либо (с глаголом)’: *желать поесть*;
2. ‘желать чего-либо (с существительным)’: *желать отдыха*.

Каждое из значений было отдельно обработано в ONTOHELPER. Разница между ними заключается в том, что в первом случае действие направлено на ‘действие или состояние’, во втором - на ‘процесс’. Для обоих глаголов в качестве процесса было указано существительное *желание* в соответствующих значениях.

Таким образом, для устранения разрывов в словосочетаниях с определительно-объектными отношениями основная работа заключалась в исправлении в онтологии концептов-существительных и обработке глаголов в онторедакторе ONTOHELPER.

2.3.1.2 Словосочетания с определительно-субъектными отношениями

В данную группу вошли такие сочетания двух именных групп, в которых главная составляющая называет действие, а второе, стоящее в родительном падеже, – производителя этого действия. При этом зависимая составляющая может называть лицо, предмет или отвлеченное понятие.

При устранении разрывов в словосочетаниях с определительно-субъектными отношениями множество таких словосочетаний было разделено на группы по способу устранения разрывов в них. Рассмотрим отдельно каждую группу.

В первую группу вошли словосочетания, встретившиеся, например, в следующих контекстах:

- «…среди ночи проснулся, не то уже выспавшись [...], не то от ровного *шума дождя по железной крыше*…» [Андреев: 1990: 168] (курсив наш. - М. Г.);
- «*Грохот дождя по крыше* стал глуше…» [Там же: 169] (курсив наш. - М. Г.);
- «… донесся *грохот экипажа по мостовой*…» [Там же: 176] (курсив наш. - М. Г.);
- «… домик затрясся от *грохота и лязга взошедшего на мост поезда*…» [Там же: 180] (курсив наш. - М. Г.) и т.д.

Проблема в случае с первым примером (*шума дождя по железной крыше*) заключалась не в возникновении разрыва, а в неверном для данного контекста синтаксическом и семантическом разборе. В приведенном отрывке речь идет о звуке (шуме), исходящем от стука капель по крыше, т.е. капли дождя, являющиеся совокупностью материальных объектов, производят шум (шумят). Словоформа *дождя* в данном случае выступает в качестве дополнения, а не в функции несогласованного определения. Для такого значения словосочетания будут верны разборы, приведенные на рис. 3 и рис.

4.

Рисунок 3 Корректный синтаксический разбор словосочетания *шум дождя*

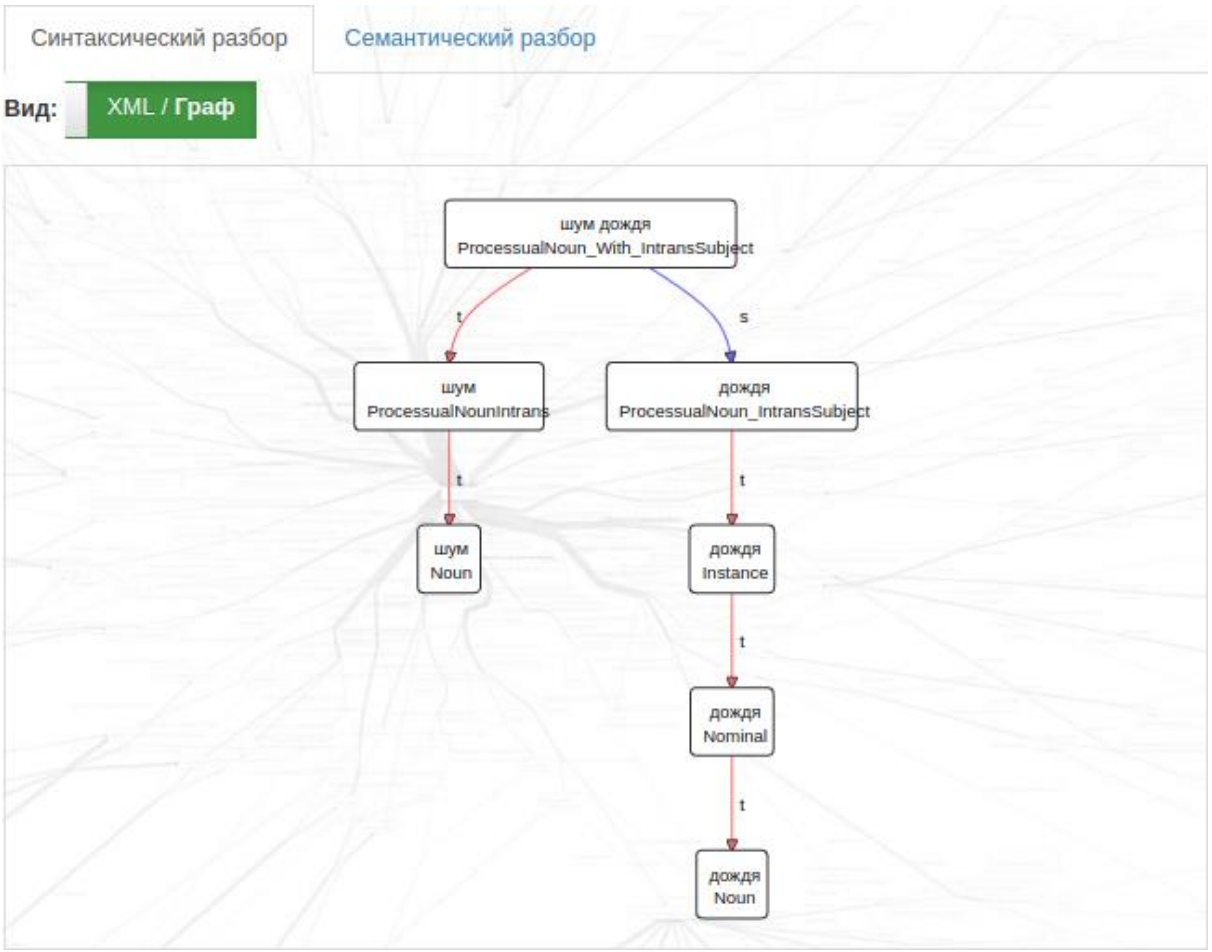


Рисунок 4 Корректный семантический разбор словосочетания *шум дождя*

При автоматическом анализе приведенного словосочетания в тексте

рассказа возникла проблема: слово *дождь* было разобрано как несогласованное определение. К этому привело отсутствие описаний значений слов *дождь* и *шум* в онтологии. Следствием этого стала семантическая интерпретация этих слов как ‘неизвестный объект или процесс’, а также семантический анализ данного словосочетания ‘неизвестный объект или процесс’ обладает ‘неизвестным объектом или процессом’, который можно интерпретировать как ‘дождь обладает шумом’. Некорректный в данном случае разбор представлен на рис. 5 и рис. 6.

Рисунок 5. Некорректный синтаксический разбор словосочетания *шум дождя*

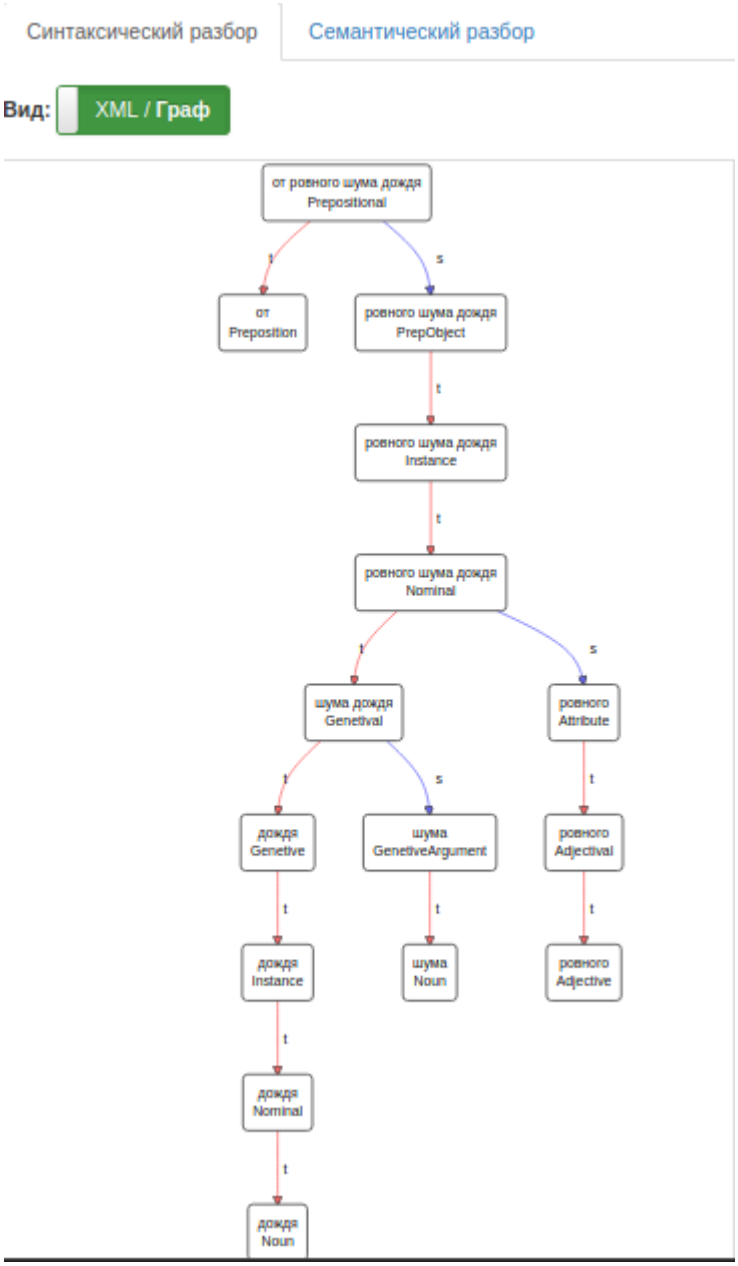


Рисунок 6. Некорректный семантический разбор словосочетания *шум дождя*

Для решения описанной проблемы в первую очередь был изменен надкласс слова *дождь* в онтологии. Предыдущий надкласс ‘обложные осадки’, автоматически добавленный в значение, не удовлетворял условиям решения задачи и был заменен. Значению слова *дождь* (‘жидкие осадки в виде капель диаметром от 0,5 до 5 мм’) был присвоен надкласс ‘совокупность капель’ с той целью, чтобы представить дождь в виде совокупности материальных объектов. Это понадобилось затем, чтобы был возможен верный разбор словосочетания *шумит дождь*: поиск в Национальном корпусе русского языка [Национальный корпус русского языка] конструкций с глаголом шуметь показал, что издавать шум могут только материальные объекты: шум колес, шум трамвая, шум воды. Даже такие примеры как *шум ветра, шум базара* на самом деле являются метонимией: шумит не ветер, а то, что он приводит в движение; не базар, а люди на базаре.

Для достижения надкласса ‘совокупность материальных объектов’ были добавлены следующие надклассы для концепта ‘совокупность капель’: ‘совокупность частиц’, ‘совокупность частей материального объекта’, являющихся соответственно надклассами концепта ‘совокупность капель’ и подклассами концепта ‘совокупность частиц’.

Вторым шагом при решении проблемы разбора *словосочетания шум дождя* стало моделирование валентностей глагола *шуметь* с помощью онторедактора ONTOHELPER. Значение глагола было описано как

‘ненаправленная неадресованная деятельность, субъектом которой является материальный объект’. Процесс осуществления этой деятельности выражается существительным шум.

Таким образом, представив дождь в качестве совокупности капель, т.е. материальных объектов, удалось связать значение этого слова со значением слова *шум*, субъектом деятельности которого является материальный объект.

Далее возникла необходимость связать словосочетания *шум дождя* и *по крыше*, т. к. при их разборе также возник разрыв. Эта проблема напрямую связана с глаголом *шуметь*, т. к. он управляет предлогами *по* и *в*, и это необходимо отразить в онтологии для установления необходимых валентностей.

При работе с корпусами и текстами в AIIRE сотрудниками проекта была собрана коллекция семантических классов глаголов, каждый из которых имеет свой набор предлогов, обеспечивающих глаголам из этих классов необходимые семантические валентности. Так, глагол *шуметь*, помимо всего прочего, относится к классу ‘осуществлять/осуществить действие или состояние по нанесению удара’ и потому способен управлять предлогами в и по. Такой класс для глагола *шуметь* был выбран исходя из того, что дождь шумит по крыше посредством ударов капель о нее (так, словосочетание *шум дождя по крыше* вполне заменяется на *стук дождя по крыше*). Поиск похожих конструкций для глагола шуметь в Национальном корпусе русского языка [Национальный корпус русского языка] выявил, что в случаях его употребления с предлогом по и обстоятельством места в качестве последнего чаще всего встречаются ‘по крыше’, ‘по деревьям’, ‘по листве’ и т. д., а субъектом действия является некое природное явление: дождь, ветер. После указания класса глаголов ‘осуществлять/осуществить действие или состояние по нанесению удара’ для значений глаголов *шуметь, грохотать, стучать* и т.д. разрыв между словосочетаниями *шум дождя* и *по крыше*, а также в подобных конструкциях, был устранен.

В грамматику также был внесен новый класс для составляющей, в

которую входят процессуальное существительное, соответствующее непереходному глаголу, с именной группой, обозначающей субъект – ProcessualNoun_With_IntransSubject. Добавление такого класса позволило избежать некорректного разбора подобной составляющей, как в примере, представленном на рис. 4.

Во вторую группу вошли словосочетания, примеры которых приведены в следующих контекстах:

- «… а снаружи неслись бодрые звуки голосов и отчаянно-звонкое *щебетанье воробьев*» [Андреев 1990: 170] (курсив наш. - М. Г.);
- «… слышались визгливые женские голоса и *плач детей*» [Там же: 174 (курсив наш. - М. Г.)];
- «… сдержанный веселый *говор людей*…» [Там же: 176] (курсив наш. - М. Г.);
- «… и *звук его голоса* отскочил от крыши …» [Там же: 178] (курсив наш. - М. Г.);
- «… радуясь *звуку своего голоса* и возможности поговорить…» [Там же: 180] (курсив наш. - М. Г.) и т.д.

Устранение разрывов в этих словосочетаниях происходило по единой схеме:

1. В онторедакторе ONTOHELPER были обработаны *глаголы звучать, щебетать, плакать, говорить* и др. Значения их были определены как деятельность (т.к. для них невозможно указать видовые пары, обозначающие завершение деятельности), т.е. для каждого глагола была указана его форма в несовершенном виде и существительное, указывающее на процесс по действию глагола. Стоит отметить, что при обработке глагола *звучать* в онторедакторе возник вопрос о значении существительного *звук*. Проблема возникла из-за конфликта толкований данного существительного с научной точки зрения (колебания воздуха, воспринимаемые слухом) и языковой картины мира (звук как процесс деятельности, описываемой глаголом звучать). Было найдено

следующее решение: в онтологию было добавлено два значения слова *звук* для каждого из толкований. Им были указаны надклассы ‘волна (форма переноса энергии, характеризующаяся периодическим изменением параметров физической среды)’ и ‘ненаправленная неадресованная деятельность звука’, соответственно. При обработке глагола *звучать* существительное *звук* было указано и в качестве процесса этого действия (в значении ‘создание высокочастотных колебаний воздуха или иной среды, воспринимаемых слухом’), и его субъекта (в значении ‘совокупность высокочастотных колебаний воздуха или иной среды, воспринимаемых слухом’).

2. В онтологию были добавлены ранее отсутствующие в ней концепты, такие как *воробей*, и дополнены уже существующие, такие как *дети*.

Как уже было сказано выше, при синтаксическом анализе используется также модуль синтаксической семантики, за счет которой существенно сокращается количество комбинаторных взрывов при разборе. Необходимость устранения проблем, связанных с семантикой и описанных выше, хоть на первый взгляд не входит в рамки работы по исследованию проблем синтаксиса, возникла в том числе и для устранения тех разрывов, которые препятствовали связыванию двух конструкций, проблемы которых были обусловлены синтаксисом. Ниже приведены некоторые примеры словосочетаний с определительно-субъектными отношениями, которые не были объединены в группы, в отличие от, например, словосочетаний типа шум дождя или звук голоса, но решение проблем, возникающих при их анализе, было однотипно.

- «… с тем приятно-жутким чувством, с каким люди встречают *проявление грозной силы природы*» [Андреев 1990: 169] (курсив наш. - М. Г.);
- «… несло свежестью и тиной, и неуловимым *запахом льда, воды и навоза*» [Там же: 169] (курсив наш. - М. Г.);
- «Звонкие голоса мельников, ловивших доски, *сияние неба и солнца*,

разноголосый крик на той стороне...» [Там же: 172] (курсив наш. - М. Г.);

- «… вздрагивая от *прикосновения холодных ветвей*…» [Там же: 177] (курсив наш. - М. Г.);
- «Под нею чуялось *присутствие живых людей*…» [Там же: 177] (курсив наш. - М. Г.) и т.д.

Проблемы разрывов в перечисленных и подобных им конструкциях решались путем добавления в онтологию новых концептов и значений и редактирования уже существующих, но имеющих не подходящее для настоящей работы описание или не утвержденных лингвистами. Ниже приведен пример описания работы над *конструкцией проявление грозной силы природы.*

Задачей при работе над этой конструкцией стало описание ее как чего-то нематериального, проявляющегося в материальном мире. Был создан класс ‘нематериальный объект или процесс’, надклассом которого стал концепт ‘объект или процесс’, а подклассами – ‘нематериальный объект’ и ‘процесс’. Далее в онторедакторе ONTOHELPER было создано значение глагола *проявляться*, описанное как ‘оказываться раскрытым’, был указан глагол *проявиться* в качестве завершения и существительное *проявление* как указывающее на процесс по этому действию. Субъектом действия был указан ‘нематериальный объект или процесс’. Так, учитывая, что в онтологии для значения слова сила ‘источник чего-нибудь, какой-нибудь деятельности, явления’ был указан надкласс ‘нематериальный объект’; и принимая во внимание все внесенные изменения, стал возможным анализ конструкции *проявление грозной силы природы* без разрывов и построение для нее единого синтаксического дерева и концептуального графа, представленных на рис. 7 и рис. 8.

Рисунок 7 Синтаксический разбор словосочетания *проявление грозной силы природы*

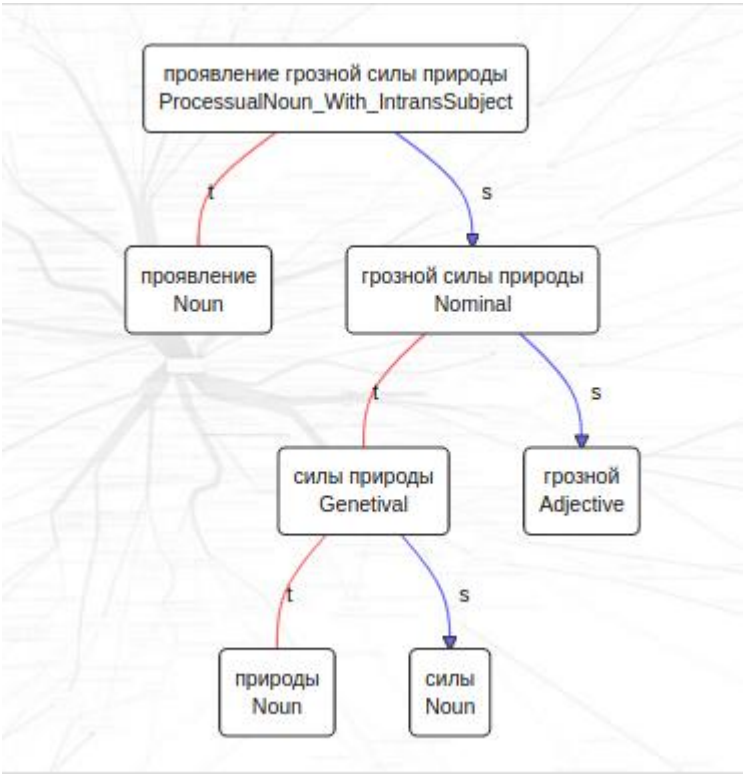


Рисунок 8 Семантический разбор словосочетания *проявление грозной силы природы*

2.3.2 Генитивные конструкции

Следующая группа сочетаний именных групп, одна из которых употреблена в родительном падеже – группа с генитивными конструкциями.

До сих пор нет устоявшегося определения генитивных конструкций. В работе [Борщев, Парти 1999] говорится, что «Семантику русской генитивной конструкции (учитель музыки) нетрудно описать в общих, не слишком четких терминах. Она задает реалию, обозначаемую опорным существительным (учитель), находящуюся в некотором отношении к другой реалии, обозначаемой генитивной именной группой (музыки). Причем выбор отношения определяется обычно опорным существительным» [Борщев, Парти 1999: 159]. А.А. Леонтьев пишет, что «Генитивные отношения (ГО) – это способ представления семантики генитива. То есть вся область значений, которые могут быть выражены генитивом дробится на более мелкие значения, которые я и называю ГО. Например, ГО это отношения «часть тела» или отчуждаемой vs неотчуждаемой принадлежности. До сих пор не существует единого списка ГО, некоторые исследователи [...] даже считают, что вследствие абстрактности семантики генитива, создать универсальный набор ГО невозможно» [Леонтьев 2005]. В работе [Леонтьев, Леонтьева 2006] приводятся следующие примеры генитивных отношений: «часть – целое» (ручка двери), «Материал – объект» (доски пола, железо крыш), «группа – представители группы» (колонна ребят, наша бригада) и другие [Леонтьев, Леонтьева 2006: 336]. Стоит отметить, что в проекте AIIRE генитивными считаются отношения любого вида принадлежности объекта или процесса объекту или процессу, а отношения между процессом и его субъектом (объектом) собственно генитивными не считаются.

Отношения – основа онтологии AIIRE. Иерархия концептуальных отношений является основным источником ограничений для лексической неоднозначности. Каждое отношение связано с классами его субъектов и объектов (в онтологии существуют отношения, названные ‘обладать субъектом’ или ‘обладать объектом’, которые помечают эти отношения), а также может иметь так называемое отражение (обратное отношение) через отношение ‘отражаться отношением’ [Dobrov 2014: 150]. Внесение в онтологию новых отношений и редактирование уже существующих стало

основной задачей при решении проблемы разрывов в конструкциях с генитивными отношениями. Всего было внесено 8 групп конструкций, для которых требовались определенные отношения. Ниже приведена характеристика каждой группы.

Материал – объект. В эту группу вошли следующие конструкции:

- «...точно над *железом крыши* опрокинули мешок с горохом…» [Андреев 1990: 168] (курсив наш. - М. Г.);
- «...лениво кружились *щепки и доски поломанных заборов*…» [Там же: 174] (курсив наш. - М. Г.) и т.д.

Отношение ‘быть сделанным из материала’, необходимое для связи подобных словосочетаний, на момент решения проблемы уже было внесено в онтологию и утверждено лингвистами. Данное отношение имеет значение ‘(о продукте) быть сделанным из материала’ и отражается отношением ‘быть материалом, из которого изготовлен продукт (о материале)’. Проблема разрыва заключалась в том, что в онтологии отсутствовали концепты для слов *щепка, доска, забор*, а также необходимое значение для слова *железо*: для данного концепта существовало только значение ‘химический элемент с атомным номером 26’. Для решения проблемы разрыва в выражение было добавлено еще одно значение ‘твердый ковкий металл’, которое по цепочке надклассов восходит к концепту ‘материал’. Также были внесены в онтологию концепты ‘щепка’, ‘доска’ и ‘забор’ (значения лексических единиц *щепка* и *доска* также восходят к концепту материал). Таким образом, разрыв в приведенных конструкциях был обусловлен не отсутствием в онтологии необходимого отношения, а невозможностью вступления концептов в это отношение из-за отсутствия или недостаточной обработки в онтологии.

Часть – целое. Эта группа – самая объемная по числу вошедших в нее словосочетаний. Ниже приведены некоторые контексты с примерами таких конструкций:

- «… расплывался четырехэтажный *корпус мельницы*…» [Андреев 1990: 169] (курсив наш. - М. Г.);

- «…кружились грязные *соломинки навоза*, смытого со двора…» [Там же: 171] (курсив наш. - М. Г.);
- «… к плечу прижималось то испуганное *личико девочки*, то восторженное и замазанное *лицо мальчугана*» [Там же: 174] (курсив наш. - М. Г.);
- «… гибкие *ветви деревьев* с разбухшими почками…» [Там же: 175] (курсив наш. - М. Г.);
- «… стукнулась о *порог домика*» [Там же: 176] (курсив наш. - М. Г.);
- «… падали с неба и поднимались со *дна реки*…» [Там же: 181] (курсив наш. - М. Г.) и т.д.

Ниже рассмотрены способы решения проблемы разрывов в таких конструкциях на примере некоторых приведенных словосочетаний.

В онтологию уже было внесено необходимое отношение ‘иметь материальную часть’ со значением ‘(о материальном объекте) иметь материальную часть’ и отражающееся отношением ‘быть частью материального объекта’. Объектом отношения ‘иметь материальную часть’ также было указано отношение ‘быть типичным представителем класса ‘иметь материальный объект (о материальном объекте)’’. В этом случае необходимо было внести новые концепты в онтологию или отредактировать уже существующие таким образом, чтобы эти концепты могли вступать в вышеупомянутое отношение ‘иметь материальную часть’.

Так, например, была решена проблема разрыва в словосочетании *корпус мельницы*: в первую очередь в выражение *мельница* было добавлено значение ‘здание, в котором расположено предприятие, ориентированное на размол зерна’. Надклассом для этого значения был установлен концепт ‘здание’, у которого в качестве объекта отношения указано отношение ‘иметь материальную часть’ ‘часть здания’. Далее был создан концепт ‘обособленная большая часть здания’, надклассом которого является концепт ‘часть здания’, а объектом отношения – ‘быть типичным представителем класса’ ‘корпус (основная часть здания)’. Таким образом, установив необходимые отношения

между определенными значениями, удалось добиться устранения разрыва в словосочетании.

Подобная работа проводилась и с остальными словосочетаниями, вошедшими в данную группу. Основная работа, как описано выше, заключалась в том, чтобы связать определенные значения слов с такими надклассами, которые имеют способность вступать в необходимые отношения (т.к., как упоминалось в главе I, параграфе 4 настоящей работы, элементы онтологии расположены иерархично и способны наследовать признаки более высоких классов).

Область пространства – субстанция. В эту группу вошли, например, следующие словосочетания:

- «…золотой *столб солнечного света*…» [Андреев 1990: 170] (курсив наш. - М. Г.);
- «… и тьму прорезал *столб электрического света*…» [Там же: 181] (курсив наш. - М. Г.);
- «… на бледном лице ее горит *отблеск далекого белого света*» [Там же: 181] (курсив наш. - М. Г.) и др.

На момент решения проблемы лексическая единица *свет* уже была описана в онтологии и имела значение ‘электромагнитное излучение, воспринимаемое человеческим глазом’. Через ряд надклассов оно восходит к концепту ‘нематериальный объект или процесс’. В онтологию были добавлены концепты для существительных *столб* и *отблеск*, имеющих значения ‘вертикальная область пространства цилиндрической формы’ и ‘блеск, отсвет на какой-либо поверхности’, соответственно. Оба этих концепта через ряд надклассов восходили к лексической единице *компонента* со значением ‘объект, являющийся частью другого объекта, противоположность самостоятельному объекту’. Значение ‘компонента’ являлось объектом отношения ‘обладать типичным представителем’ ‘часть нематериального объекта’ (такое отношение также было указано и для материального объекта). Таким образом, учитывая, что концепт ‘свет’ наследует признаки

нематериального объекта, концепты вступили в необходимые отношения, и разрыв был устранен.

Область пространства объекта. В эту группу вошли словосочетания из следующих контекстов:

- «… она поднесла к глазам *конец платка*…» [Андреев 1990: 181] (курсив наш. - М. Г.);
- «Со всех *концов темного горизонта* лились медные голоса…» [Там же: 181] (курсив наш. - М. Г.);
- «И во всех *концах горизонта* начали зажигаться красные и голубые огни…» [Там же: 181] (курсив наш. - М. Г.) и др.

При работе со словосочетаниями из настоящей группы следовало обратить внимание на то, что объект, выраженный именной группой в родительном падеже, может быть как материальным, так и нематериальным. Принимая во внимание этот факт, концепт ‘конец’ был обработан максимально абстрактно, чтобы вступать в отношения с концептами, имеющих признаки и материальных, и нематериальных объектов. Ему было присвоено значение ‘предел, граница, край какого-то объекта, а также его часть, примыкающая к этому пределу’ и надкласс ‘часть’ являющийся синонимом концепта ‘компонента’, описанного выше. Концепты ‘горизонт’ и ‘платок’ были внесены в онтологию и восходили через ряд надклассов к концептам ‘нематериальный объект’ и ‘материальный объект’, соответственно; следовательно, могли вступать с концептом ‘конец’ в необходимые отношения.

Стоит отметить, что подобным образом – внесением и редактированием в онтологии новых концептов и присваиванием им таких надклассов, чтобы они могли вступать в необходимые отношения – были также обработаны такие группы конструкций как *временной период – процесс* (например: «… за шесть дней непрерывного лежания...» [Андреев 1990: 169]) (курсив наш. - М. Г.) и *совокупность людей – человек* (например: «… он не мог принадлежать к старшему поколению стрельцов» [Андреев 1990: 174] (курсив наш. - М. Г.)).

Ниже будут рассмотрены такие генитивные конструкции, для решения проблемы разрывов в которых были внесены новые отношения в онтологию.

Учебник – предмет. Для устранения разрыва в словосочетании «попробовал читать самоучитель французского языка» [Андреев 1990: 173] (курсив наш. - М. Г.) необходимо было внести в онтологию новые отношения и установить их между определенными концептами. Задача была выполнена в несколько этапов:

1. В выражение язык было добавлено новое значение ‘учебная дисциплина’, подклассами которого являются концепты ‘английский язык’, ‘французский язык’ и ‘русский язык’.
2. В выражение учебник было добавлено новое значение ‘книга или другой носитель информации с систематическим изложением знаний в определенной области для занятий как в образовательной системе, так и в ходе самостоятельного обучения’.
3. Было создано два концепта-отношения: ‘обладать учебником’ со значением ‘(о предмете) обладать учебником’ и отражение этого отношения быть учебником со значением ‘(об учебнике) быть учебником предмета’.
4. Концепт ‘предмет’, который является надклассом значения ‘учебная дисциплина’ лексической единицы язык, был указан в качестве объекта отношения ‘обладать учебником’ ‘учебник’; а концепт ‘учебник’ – в качестве объекта отношения ‘быть учебником предмета’ ‘предмет’.

Таким образом, создав в онтологии новые отношения и обозначив концепты, которые могут в них вступать, удалось добиться корректного семантического разбора и построения для словосочетания единого синтаксического дерева.

Звуки – период. Сюда вошли такие словосочетания как «… и скоро все тихие *звуки ночи* утонули...» [Андреев 1990: 181] (курсив наш. - М. Г.). Проблема разрыва была решена следующим образом:

1. Для лексической единицы *звук*, помимо упомянутых значений, было

создано два значения: ‘создание высокочастотных колебаний воздуха или иной среды, воспринимаемых слухом’ с надклассом ‘ненаправленная неадресованная деятельность звука’ (указание такого значения было бы корректным при разборе словосочетания звук голоса) и ‘совокупность высокочастотных колебаний воздуха или иной среды, воспринимаемых слухом’ с надклассом ‘волна’. Второе значение по цепочке надклассов восходит к концепту ‘процесс’.

2. Были созданы концепты-отношения ‘обладать характерным процессом’ со значением ‘(о регулярно повторяющемся временном периоде) обладать характерным процессом’ и его отражение ‘принадлежать регулярно повторяющемуся временному периоду’ со значением ‘о характерном процессе’.
3. Был отредактирован существующий в онтологии концепт ‘регулярно повторяющийся временной период’ (к этому концепту по цепочке надклассов восходит значение концепта ‘ночь’): в качестве его объекта отношения был указан концепт ‘обладать характерным процессом’ ‘процесс’.

Таким образом, значение слова *ночь* как регулярно повторяющийся временной период, благодаря указанному отношению может обладать значением слова *звук* как характерным для него процессом.

Из приведенного выше описания работы с генитивными конструкциями можно сделать вывод, что при устранении в них проблем, возникающих при автоматическом синтаксическом анализе, существенное значение играет семантическая составляющая: без установления определенных отношений между словоформами не может быть выбрана корректная версия дерева синтаксического анализа. По этой причине работа с онтологией занимает важное место при создании синтаксически размеченного корпуса художественных текстов.

Представляется важным также отметить, что в ходе работы по устранению разрывов в описанных в настоящем параграфе конструкциях

время от времени возникали «локальные» комбинаторные взрывы, не отраженные в корпус-менеджере, но замеченные при проверке конструкции в лингвистическом процессоре. Такие взрывы возникали из-за внесения в онтологию новых концептов и отношений, которые могли нарушать установленную ранее иерархию некоторых концептов, приводя к большому количеству семантических разборов. В таком случае взрывы устранялись непосредственно при работе с конструкцией путем описанных выше изменений в онтологии.

2.4 Деепричастные обороты

В.В. Виноградов писал, что «обособленные второстепенные члены предложения представляют собой одну из разновидностей выделяемых в предложении интонационно-смысловых отрезков, но обязательно таких, которые всегда представляют собой грамматически связанное целое» [Виноградов, Истрина 1960: 641]. Среди обособленных обстоятельств Виноградов выделяет «обособленные обстоятельства, выраженные деепричастием – одиночным или с относящимися к нему словами» [Там же: 650]. Деепричастие с относящимися к нему словами также называют деепричастным оборотом.

При каждом деепричастном обороте при разборе рассказов встречались разрывы. Это было связано с тем, что конструкции для обособленных деепричастных оборотов в грамматике AIIRE отсутствовали: такие конструкции не характерны для текстов новостных сообщений, зато они часто встречаются в художественных текстах. Изначально в грамматике был прописан класс Gerundial – для необсобленного деепричастного оборота, а также класс VerbialWithGer для глагольной группы с необособленным деепричастным оборотом. Очевидно, что в случае, когда оборот был выделен запятыми, возникал разрыв, т.к. в классе Gerundial не предполагалось наличие запятых (они не были указаны в качестве аргументов в массиве

argument_classes). Стоит также отметить, что для класса Gerundial в качестве аргумента была указана закрывающая кавычка (предполагается, что это была либо опечатка при составлении правил для составляющей, либо один из текстов в корпусе в какой-то момент потребовал написания именно такого правила). В связи с этим возникла необходимость добавления в грамматику конструкции обособленного запятыми обстоятельства, выраженного деепричастным оборотом.

В первую очередь в грамматику были добавлены классы ClosedGerundial – конструкция, описывающая оборот, обособленный знаками препинания с двух сторон, – и OpenGerudial – для открытого, то есть не имеющего знака препинания справа обстоятельства, выраженного деепричастным оборотом. OpenGerudial является главной составляющей для ClosedGerundial, а для OpenGerudial, в свою очередь, являются главными классы Verbial (описывающие глагольную группу), Verbial_TimeBefore (описывающая глагольную группу с обстоятельством времени предшествования) и Verbial_TimeAfter (описывающая глагольную группу с обстоятельством времени последования). Были также описаны правила, учитывающие постановку знаков препинания.

При проверке разбора с учетом вышеописанных классов в грамматику, однако, возникла проблема. Изначально предполагалось, что деепричастный оборот должен связаться с глагольной группой при помощи непроективного порядка. Это значит, что в предложении *Набросив пальто на голову, он обошел…* [Андреев 1990: 169] сказуемое, выраженное глагольной группой *обошел* может связаться с деепричастным оборотом *Набросив пальто на голову*, минуя подлежащее *он*. Непроективный порядок действительно учтен в грамматике AIIRE, и для него прописаны правила. Возникшая проблема, однако, заключалась в том, что класс VerbalWithGer обладает техническим ограничением, предполагающим недопущение эллипсиса как главного, так и зависимого класса, а значит, недопущение также непроективного порядка слов. Без этого ограничения конструкция VerbalWithGer встречалась бы во

всех разборах.

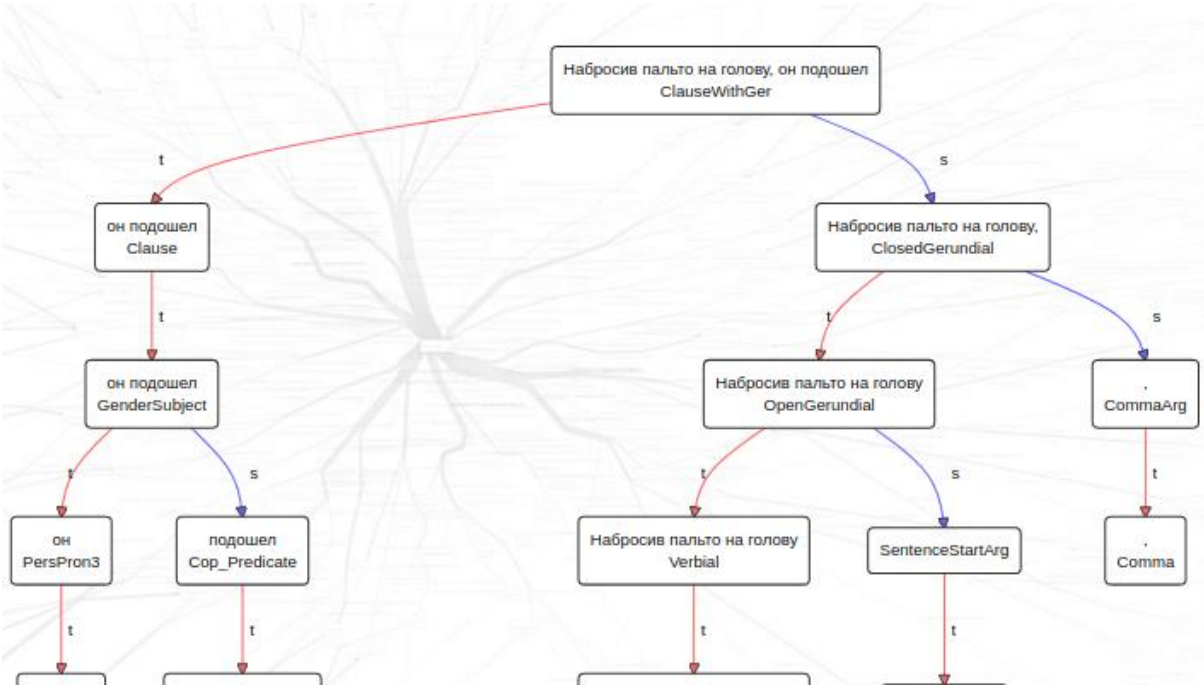
При решении этой проблемы разумно было бы предположить, что деепричастный оборот относится не исключительно к глагольной группе, но ко всему предложению в целом, как это происходит с детерминирующими обстоятельствами. Стоит, однако, отметить, что здесь некорректно использовать термин ‘детерминирующее обстоятельство’, т.к., вводя этот термин, Н. Ю. Шведова подразумевала под ним наречия или предложно-падежные сочетания, «структурно совпадающие с именным компонентом простого двучленного словосочетания с необязательной связью» [Шведова 1964: 81]. Позже Г.А. Золотова предложила использовать термин ситуант для «свободных синтаксических форм, распространяющих предикативный минимум» [Золотова 1973: 105], однако в ее работе [Золотова 1973] речь также идет о предложных и предложно-падежных формах; деепричастные обороты при этом не упоминаются. Тем не менее, для разрешения описанной выше проблемы становится необходимым внести в грамматику новый класс, позволяющий связывать обособленный деепричастный оборот не непосредственно с глагольной группой, а со всем предложением.

Таким образом, в грамматику был добавлен класс ClauseWithGer, позволяющий связывать деепричастные обороты с остальным предложением. Стоит отметить, что у данного класса в argument_classes входит только класс ClosedGerundial, т. к. использование деепричастного оборота, примыкающего ко всему предложению, предполагает обособленность, то есть наличие знаков препинания с двух сторон. Класс ClauseWithGer был добавлен в качестве главного класса в те классы, которые имели при себе класс Clause (описывающий клаузу) в массиве, содержащем перечень главных классов. Это было сделано исходя из предположения, что к каждому случаю, описанному в Clause, может присоединяться деепричастный оборот.

Таким образом, после внесения в грамматику класса ClauseWithGer и связанных с ним корректировок, примеры типа *Набросив пальто на голову, он обошел…* разбираются без разрывов, т. к. деепричастный оборот

присоединяется к целому предложению, а не к исключительно глагольной группе. В итоге, верный разбор для данного словосочетания приведен на рис. 9.

Рисунок 9 Синтаксический разбор словосочетания *Набросив пальто на голову, он подошел*



2.6 Критерии оценки синтаксического анализа и качества разметки

Для того, чтобы оценить полноту синтаксической разметки, в общем случае необходимо наличие золотого стандарта, с которым происходило бы сравнение результатов анализа. На этапе создания синтаксически размеченного корпуса рассказов Л. Андреева такого стандарта не существует (т.к. корпус и есть золотой стандарт – его разработка все еще продолжается). По этой причине в настоящий момент оценивать полноту точность разметки корпуса можно лишь по косвенным признакам.

Полнота синтаксической разметки – отношение количества представленных в разметке синтаксических единиц (конструкций) к общему числу синтаксических единиц в корпусе. Вторая величина в настоящий

момент неизвестна в силу отсутствия золотого стандарта, однако известно количество разрывов – случаев, когда корректная синтаксическая конструкция отсутствует в разметке. Соответственно, уменьшение числа разрывов говорит об уменьшении количества таких случаев, т.е. об увеличении полноты разметки.

Точность разметки также косвенно связана с количеством версий разбора: обычно при наличии комбинаторного взрыва (слишком большого количества версий) существенная их часть некорректна. Следовательно, в общем случае, чем больше случаев комбинаторных взрывов, тем ниже точность разметки. В настоящей работе наибольшие по объему комбинаторные взрывы были вызваны следующими причинами:

1. Лексические единицы типа «Дур-р-а!» или «р-рука» («На реке» (1900)) вызывают ситуации, когда возникает несколько нераспознанных элементов подряд: «р» анализируется как существительное, обозначающее букву, во всех падежах – это приводит к череде комбинаторных взрывов. Такие проблемы автоматического анализа текста в настоящий момент не решены;
2. Эллипсис существительного в однородных именных группах, например: «Оттуда несло ароматным теплом, *ласкавшим горло и щекотавшим в носу*» [Андреев 1990: 171] (*курсив наш – М. Г.*). В таком случае ряд однородных причастных оборотов распознается как имеющий эллипсис существительного. Это приводит к тому, что неизвестен род и одушевленность именной группы – предполагается, что он может быть любым, и возникает комбинаторный взрыв.

Данные случаи комбинаторных взрывов, хотя и являются не решенными проблемами, тем не менее, не могут быть решены на данном этапе исследования именно потому, что сами по себе не содержат явно некорректных с точки зрения формальной грамматики версий, т.е. не снижают общую точность; что же касается некорректных версий разметки в исследованных конструкциях, которые влияли на точность, то все они были устранены в ходе исследования.

2.6 Выводы

1. В корпус, собранный для проведения настоящего исследования, вошли рассказы 1900 года, т.к. представляется, что этот материал репрезентативен и синтаксические конструкции, характерные для этого периода также встречаются в более ранних и более поздних текстах Л. Андреева. Объем корпуса составил 36093 словоупотреблений. Отбор текстов по жанру и году осуществлялся с опорой на собрание сочинений Л. Андреева, изданное в 1900-1996 годах.
2. Синтаксические конструкции, при автоматическом анализе которых возникали проблемы, выбирались вручную в зависимости от частоты появления в текстах и систематичности появления разрыва. В настоящей работе проанализировано 4 группы (10 подгрупп) таких конструкций, и в дальнейшем предполагается работа и над другими группами.
3. Проблемы возникновения разрывов между именными группами, одна из которых употреблена в родительном падеже, решались несколькими способами:
    a. добавление в онтологию новых концептов, в том числе с использованием онтологического редактора ONTOHELPER;
    b. редактирование уже существующих концептов, внесение в них новых значений и установление отношений между концептами;
    c. создание новых отношений в онтологии.

4. Проблема возникновения разрыва между определяемым словом и обособленным приложением и синтаксическими конструкциями с деепричастными оборотами была решена путем добавления в грамматику новых классов, описывающих составляющие и зависимости между ними.
5. Количество разрывов в анализе рассказа «На реке» (1900) по окончании работы над формализацией грамматики и синтаксической неоднозначности было подвергнуто сравнению с количеством разрывов в начале работы. Обнаружилось, что в процессе работы количество разрывов сократилось с 4446 до 3285.

Заключение

Исследование проблем автоматического синтаксического анализа художественных текстов на материале рассказов Л. Андреева позволило сделать следующие выводы:

1. Проблемы при автоматическом синтаксическом анализе художественных текстов с помощью инструментов проекта AIIRE возникали в основном в конструкциях, характерных для текстов художественного стиля (деепричастные обороты, обособленные приложения, сочетания именных конструкций, одна из которых употреблена в родительном падеже). Это обусловлено тем, что художественные тексты ранее не подвергались анализу в данном проекте.
2. Решение проблем, возникающих при автоматическом синтаксическом анализе, осуществлялось следующими способами:
    a. добавление в грамматику новых классов непосредственных составляющих (в ходе работы было добавлено 8 новых классов) и редактирование существующих;
    b. работа в онтологии: создание новых выражений и значений, установление между ними необходимых связей;
    c. работа в онтологии: создание новых отношений и добавление их в существующую иерархию отношений в онтологии.
3. Полнота разметки в ходе исследования увеличилась на 26%, однако число комбинаторных взрывов увеличилось на единицу. Стоит отметить, что сократить число комбинаторных взрывов (а значит, увеличить точность разметки) в рамках настоящего исследования не удалось, т.к. часть из них иллюстрируют в настоящий момент не решенную проблему нераспознанных единиц типа ‘р-р-рука’; а другая часть взрывов возникает из-за проблемы эллиптированного существительного в ряду однородных определений, общей для

художественных и не художественных текстов. Тем не менее, комбинаторные взрывы, возникающие «локально» при работе над конкретной конструкцией в ходе устранения в ней разрыва, устранялись параллельно путем редактирования значений лексических единиц в онтологии. Таким образом, при незначительном снижении общей точности синтаксической разметки, точность синтаксической разметки исследованных конструкций оставалась близкой к единице в течение всего исследования.

4. При корректном семантическом анализе (когда онтология проработана и при анализе не возникает проблем на уровне семантики) в большинстве случаев строится корректное синтаксическое дерево (в случае, если все НС его структуры описаны в грамматике). Изначально в настоящем исследовании не стояла задача устранения разрывов, но в ходе работы стало очевидно, что часто этого достаточно, чтобы связать те части дерева, которые были построены для соседних конструкций, но не связывались между собой из-за неполного семантического анализа. Таким образом, чтобы избежать лишних версий разбора на синтаксическом уровне, часто достаточно обеспечить корректное связывание единиц на уровне семантики.

Список использованной литературы

1. Андреев Л. Н. Полное собрание сочинений в 23 томах. Том 01. Рассказы. 1892-1899. – М.: Наука, 2007. — 812 с.
2. Андреев Л. Н. Полное собрание сочинений в 23 томах. Том 05. Повести и рассказы. 1906-1907. – М.: Наука, 2012. — 818 с.
3. Андреев Л. Н. Полное собрание сочинений в 23 томах. Том 06. Рассказы и повести. 1908. – М.: Наука, 2013. — 780 с.
4. Андреев Л. Н. Собрание сочинений. В 6-ти т. Т. 1. Рассказы 1898-1903 гг. – М.: Худож. Лит., 1990. – 639 с.
5. Андреев Л. Н. Собрание сочинений. В 6-ти т. Т. 2. Рассказы; Пьесы 1904-1907. – М.: Худож. Лит., 1990. – 559 с.
6. Андреев Л. Н. Собрание сочинений. В 6-ти т. Т. 3. Рассказы; Пьесы 1908-1910. – М.: Худож. Лит., 1994. – 655 с.
7. Андреев Л. Н. Собрание сочинений. В 6-ти т. Т. 4. Рассказы. Сашка Жегулев. 1910-1913 гг. – М.: Худож. Лит., 1994. – 658 с.
8. Андреев Л. Н. Собрание сочинений. В 6-ти т. Т. 5. Рассказы; Пьесы 1914-1915; Сатирические миниатюры для сцены 1908-1916. – М.: Худож. Лит., 1995. – 511 с.
9. Андреева И. В. Ядро семантико-синтаксического поля бытийности: модели с собственно-бытийной семантикой (на материале прозы Л. Андреева) //Вестник Нижегородского университета им. НИ Лобачевского. – 2015. – №. 4.
10. Андрющенко В. М. Автоматическая обработка текста // Ярцева В. Н., Арутюнова Н. Д. (ред.). Большой энциклопедический словарь: Языкознание. – Большая Российская энциклопедия, 1998. – 685 с.: ил.
11. Апресян Ю. Д. Непосредственно составляющих метод // Ярцева В. Н., Арутюнова Н. Д. (ред.). Большой энциклопедический словарь: Языкознание. – Большая Российская энциклопедия, 1998. – 685 с.: ил.
12. Блумфилд Л. Язык. / Перевод с английского Е.С. Кубряковой и В.П. Мурат. Комментарий Е.С. Кубряковой. Под редакцией и с предисловием

М.М. Гухман — М.: Прогресс, 1968

13. Бондарева Н. А. Главная тема творчества Леонида Андреева //Вестник Костромского государственного университета. – 2009. – Т. 15. – №. 3.
14. Борщев В. Б., Парти Б. Х. Семантика генитивной конструкции: разные подходы к формализации //Типология и теория языка: От описания к объяснению: К. – 1999. – С. 159-172.
15. Борщев В. Б., Хомяков М. В. Клубные системы (формальный аппарат для описания сложных систем) //Научно-техническая информация. Сер. – 1976.
16. Бочаров В. В., Митренина О. В. Компьютерная морфология // Прикладная и компьютерная лингвистика/ Под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо. – М.: ЛЕНАНД, 2016. – 320 с.
17. Виноградов В. В., Истрина Е. С. Академическая грамматика русского языка в 2-х томах. Т. 2, Изд. Академии Наук СССР // Москва. – 1960.
18. Вороновская И. А. Непроективный порядок слов как средство стилизации устной речи //Филологические этюды: сб. науч. ст. молодых ученых. –Саратов, 2012. Вып. 15: в 2 кн. Кн. 2. –348 с. – 2012. – С. 193-200.
19. Гладкий А. В. Синтаксические структуры естественного языка в автоматизированных системах общения. – 1985. – 144 с.
20. Гребенников А. О. Частотный словарь рассказов Л.Н. Андреева / Под ред. Г. Я. Мартыненко. – Изд-во С.-Петерб. Ун-та, 2003.
21. Гроховский П. Л., Добров А. В., Доброва А. Е., Сомс Н. Л. Корпус-менеджер для морфосинтаксической разметки: опыт разработки корпуса тибетских грамматических сочинений // Труды международной конференции «Корпусная лингвистика-2017». – Спб.: Изд-во С.-Петерб. Ун-та, 2017. – 340 с.
22. Добров А. В. Автоматическая рубрикация новостных сообщений средствами синтаксической семантики. СПб, 2014: дис. – Диссертация на соискание ученой степени к. ф. н. СПб, 2014.

23. Добров А. В. Компьютерный семантико-синтаксический анализ языковых обозначений действий или деятельности органов государственной власти //Структурная и прикладная лингвистика. – 2015. – №. 11. – С. 111-122.
24. Добров А. В. Компьютерный синтаксис // Прикладная и компьютерная лингвистика/ Под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо. – М.: ЛЕНАНД, 2016. – 320 с.
25. Дубовик А. Р. Автоматическое определение стилистической принадлежности текстов по их статистическим параметрам // Материалы научной конференции "Интернет и современное общество". – 2017. – №. 1. – С. 29-45.
26. Зализняк А. А. Грамматический словарь русского языка: Словоизменение. М.: «АСТ-ПРЕСС», 2008. – 794 с.
27. Захаров В. П. Корпуса русского языка // Труды Института русского языка им. В.В. Виноградова. – 2015. – №. 6. – С. 20-65.
28. Захаров В. П. Корпусная лингвистика // Прикладная и компьютерная лингвистика/ Под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо. – М.: ЛЕНАНД, 2016. – 320 с.
29. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник для студентов гуманитарных вузов. – Иркутск: ИГЛУ, 2011. – 161 с.
30. Золотова Г. А. Очерк функционального синтаксиса русского языка. – 1973.
31. Иезуитова Л. А. Творчество Леонида Андреева: 1892-1906. – Изд-во Ленинградского университета, 1976.
32. Илюхин Ю. А., Сенин А. А. Леонид Андреев [Электронный ресурс]. URL: http://andreev.org.ru/biblio/rasskazi.html (дата обращения: 12.05.2018)
33. Исаева Е. В. Особенности повествования в рассказе Леонида Андреева «Цветок под ногою» // Творчество Леонида Андреева: современный взгляд. –Орел, ПФ «Картуш. – 2011. – С. 38-43.

34. Копотев М. В., Гурин Г. Б. Принципы синтаксической разметки хельсинского аннотированного корпуса русских текстов ХАНКО // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» – 2006. – С. 280-285.
35. Копотев М. В., Мустайоки А. Принципы создания Хельсинкского аннотированного корпуса русских текстов (ХАНКО) в сети Интернет //Научно-техническая информация. Сер. – 2003. – Т. 2. – С. 33-36.
36. Корпус рассказов Л. Андреева [Электронный ресурс] [2017-2018]. URL: http://aiire.org/corman/index.html?corpora_id=61&page=1&view=docs_list (дата обращения: 23.05.2018)
37. Кручинина И. Н. Приложение // Ярцева В. Н., Арутюнова Н. Д. (ред.). Большой энциклопедический словарь: Языкознание. – Большая Российская энциклопедия, 1998. – 685 с.: ил.
38. Кузнецов С. А. Глагольное изменение и формообразование в современном русском языке: дис. … докт. фил. наук: 10.02.01. СПб., 2000. – 314 с.
39. Леонтьев А. П., Леонтьева А. Л. Еще раз к вопросу о семантике генитивных отношений //Труды Международного семинара Диалог. – 2006. – С. 335-341.
40. Леонтьев А.П. Генитивные отношения в перспективе конструкций с внешним посессором в русском языке [Электронный ресурс] 2005. URL: http://www.dialog-21.ru/media/2399/leontieva.pdf (дата обращения: 19.05.2018)
41. Лукин Д. С. Война в жизни и творчестве Леонида Андреева //Труды Санкт-Петербургского государственного института культуры. – 2017. – Т. 215.
42. Маслов Ю. С. Глагол//Ярцева В. Н., Арутюнова Н. Д. (ред.). Большой энциклопедический словарь: Языкознание. – Большая Российская энциклопедия, 1998. – 685 с.: ил.
43. Машинный фонд русского языка [Электронный ресурс]. URL:

http://cfrl.ruslang.ru/ (дата обращения: 17.05.2018)

44. Митренина О. В. Проблемы неоднозначности синтаксического анализа. Автореферат диссертации на соискание ученой степени кандидата филологических наук/Санкт-Петербург, 2005.
45. Михеичева Е. А. Творчество Леонида Андреева: особенности психологизма и жанровые модификации //М.: Моск. пед. Ун-т. – 1995.
46. Национальный корпус русского языка [Электронный ресурс] [2003-2018]. URL: http://www.ruscorpora.ru/ (дата обращения: 17.05.2018)
47. Недолужко А. и др. Синтаксически аннотированный корпус чешского языка // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог. – 2008. – №. 7. – С. 400-406.
48. Попов А. М., Протопопова Е. В., Букия Г. Т. Ещё раз о способах снятия структурной омонимии: выбор единственной структуры в парсере Hurma // Материалы научной конференции" Интернет и современное общество". – 2016. – С. 65-73.
49. Протасов С . В. Обучение с нуля грамматики связей русского языка. //Десятая национальная конференция по искусственному интеллекту с международным участием., КИИ-2006. – с. 515-524.
50. Севбо И. П. Структура связного текста и автоматизаия реферирования / И.П. Севбо – Москва, 1969 – 135 с.
51. Сервис распознавания текста онлайн [Электронный ресурс] [2008-2018]. URL: https://finereaderonline.com/ru-ru (дата обращения 15.05.2018)
52. Синтаксически аннотированные корпуса – NLPub [Электронный ресурс] 2012. URL: https://nlpub.ru/Синтаксически_аннотированные_корпуса (дата обращения: 17.05.2018)
53. Синтаксически размеченный корпус русского языка: информация для пользователей. Национальный корпус русского языка [2003-2018] [Электронный ресурс] // URL: http://www.ruscorpora.ru/instruction-syntax.html (дата обращения 12.05.2018)

54. Слюсарь Н. А., Самойлова М. В. Частотности различных грамматических характеристик и окончаний у существительных русского языка. 2015 [Электронный ресурс] URL: http://www.dialog-21.ru/digests/dialog2015/materials/pdf/SlioussarNASamoilovaMV.pdf (дата обращения 17.05.2018)
55. Сомс Н. Л., Добров А. В., Доброва А. Е. Использование средств лингвистической обработки текстов в системе мониторинга информационных ресурсов по пользовательским предпочтениям //Материалы научной конференции" Интернет и современное общество". – 2014. – С. 149-158.
56. Теньер Л. Основы структурного синтаксиса: Пер. с фр. – Прогресс, 1988. – 656 с.
57. Тестелец Я. Г. Введение в общий синтаксис. – Федеральное государственное бюджетное образовательное учреждение высшего образования" Российский государственный гуманитарный университет", 2001. – 798 с.
58. Тестелец Я. Г. Эллипсис в русском языке: теоретический и описательный подходы //Москва. – 2011.
59. Хомский Н. Логические основы лингвистической теории //Новое в лингвистике. – 1965. – Т. 4. – С. 465-575.
60. Цейтин Г. С. Программирование на ассоциативных сетях //ЭВМ в проектировании и производстве. Л. – 1985. – №. 2.
61. Шведова Н. Ю. Детерминирующий объект и детерминирующее обстоятельство как самостоятельные распространители предложения //Вопросы языкознания. – 1964. – №. 6. – С. 77-93
62. AIIRE – О проекте [Электронный ресурс] [2015-2018]. URL: http://aiire.org/aiire.php (дата обращения: 17.05.2018)
63. AIIRE [Электронный ресурс]. URL: http://aiire.org 2015-2018 (дата обращения: 17.05.2017)

64. Dobrov A. V. Semantic and ontological relations in AIIRE natural language processor //Computational Models for Business and Engineering Domains. Rzeszow-Sofia: ITHEA. – 2014. – С. 147-157.

65. Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge: MIT Press, 1988

66. Hajic J. et al. Prague Dependency Treebank v2. 0. CDROM. Linguistic Data Consortium, Cat. LDC2006T01. Philadelphia. – ISBN 1-58563-370-4. URL: https://ufal.mff.cuni.cz/pdt2.0/, 2006. (дата обращения 12.05.2018)

67. Hudson R. A. Word grammar. – Oxford : Blackwell, 1984.

68. Knowledge base [Электронный ресурс] / Wikipedia, The Free Encyclopedia. Retrieved 21:23, July 3, 2014, URL: http://en.wikipedia.org/w/index.php?title=Knowledge_base&oldid=612589599 (дата обращения: 12.05.2018)

69. Krippendorff K. Combinatorial Explosion. [Электронный ресурс] / Web Dictionary of Cybernetics and Systems. PRINCIPIA CYBERNETICA WEB [2002-2012] Retrieved 29 November 2010. URL: http://cleamc11.vub.ac.be/ASC/COMBIN_EXPLO.html (дата обращения: 12.05.2018)

70. Lib.Ru/Классика: Андреев Леонид Николаевич: Собрание сочинений [Электронный ресурс]. URL: http://az.lib.ru/a/andreew_l_n/ (дата обращения: 17.05.2018)

71. Link Grammar [Электронный ресурс]. URL: http://www.link.cs.cmu.edu/link/index.html (дата обращения: 17.05.2018)

72. Link Grammar for Russian [Электронный ресурс]. URL: http://sz.ru/parser/ (дата обращения: 17.05.2018)

73. Marcus M., et al. Treebank-3 LDC99T42. Web Download. Philadelphia: Linguistic Data Consortium, 1999.

74. ONTOHELPER [Электронный ресурс]. URL: http://aiire.org/ontohelper/index.php (дата обращения: 22.05.2018)

75. Ross J. R. Constraints on variables in syntax. PhD dissertation. MIT, 1967. [Ross J. R. Infinite Syntax! Ablex: Norwood, 1986]

76. Russian Syntax Tree Bank [Электронный ресурс]. URL: http://otipl.philol.msu.ru/~soiza/testsynt/files/info.htm (дата обращения 12.05.2018)

77. Taylor A., Marcus M., Santorini B. The Penn treebank: an overview //Treebanks. – Springer, Dordrecht, 2003. – С. 5-22.

78. Tomita M. An efficient augmented-context-free parsing algorithm //Computational linguistics. – 1987. – Т. 13. – №. 1-2. – С. 31-46.

Приложение А. Классы составляющих, используемые в грамматике AIIRE

В настоящем приложении приведены классы, добавленные в грамматику AIIRE при проведении настоящего исследования.

Классы для описания обособленного приложения

```
class InstanceWithDetachedAdjunct (SyNode):
    """
    Обособленное приложение единственного числа, согласующееся с существительным в роде, числе, падеже.
    Примеры:
    Алексей Степанович, машинист при Буковской мельнице,
    Аня, девочка пяти лет,
    """
    head_classes = ['Instance']
    modifier_classes = ['DetachedAdjunct']
    modifier_can_be_ellipsed = 0
    head_can_be_ellipsed = 0
    specifier_order = Orders.Right()
    omit_grammemes = ['isName', 'nameRole']

class DetachedAdjunct (Modificational_Node):
    """
    Приложение в единственном числе, обособленное запятыми.
    Примеры:
    , машинист при Буковской мельнице,
    , девочка пяти лет,
    """
    head_classes = ['OpenDetachedAdjunct']
    argument_classes = [
            'CommaArg',
            'IsolatedPhraseEndArg',
            'SentenceEndArg',
            'TextEndArg']
    argument_can_be_ellipsed = 0
    head_can_be_ellipsed = 0
    specifier_order = Orders.Right()

class OpenDetachedAdjunct (SyNode):
    """
    Необособленное приложение, имеющее при себе только левую запятую.
    Примеры:
    , машинист при Буковской мельнице
    """
    head_classes = ['Adjunct']
    argument_classes = [
            'CommaArg',
            'IsolatedPhraseStartArg',
            'SentenceStartArg']
    argument_can_be_ellipsed = 0
```

```
    head_can_be_ellipsed = 0
    specifier_order = Orders.Left()


class Adjunct (SyNode):
    """
    Необособленное приложение.
    Примеры:
    Алексей Степанович, машинист при Буковской мельнице, →  машинист
    Аня, девочка пяти лет, → девочка
    """
    head_classes = ['Instance']
    head_can_be_ellipsed = 0

class Instance (SyNode):

    """
    Денотативная именная группа единственного числа (сигнификативная именная группа единственного числа / с предложным определением / обозначение человека по должности в том же роде / обозначение человека-женщины по должности в мужском роде / географическое наименование по классу объекта / географическое наименование / денотативная именная группа единственного числа с опеределительным придаточным предложением / группа числительного, согласованного с единственным числом / группа с неситуативным оператором  / наименование единственного числа в кавычках в роли денотативной именной группы денотативная именная группа единственного числа в кавычках / денотативная именная группа единственного числа с обособленным определением, выраженным причастным оборотом / день месяца + указательное местоимение / числовой указатель)
    Примеры:
    этот хороший работник --- InstancePointer + Nominal
    этот мальчик в шляпе --- InstancePointer + PrepModNominal
    этот президент Путин --- InstancePointer + PersonByTitle
    эта доктор Иванова --- InstancePointer + PersonByTitle_WomanMasc
    эта страна Россия --- InstancePointer + GeoObjectByClass
    этой республике Татарстан --- InstancePointer + GeoObjectByClassGoverned
    такая Россия --- InstancePointer + GeoName
    этот президент, который подписывает указы, --- InstancePointer + AttrClauseModInstance
    этот двадцать один человек --- InstancePointer + SingularNumeralGroup
    этот только президент --- InstancePointer + NonSitReflInstanceOperCall
    этот "президент" --- InstancePointer + QuotedInstance
    этот "Евгений онегин" --- InstancePointer + InstTitle
    этот президент, подписывающий указы  --- InstancePointer + IsolatedPart_Mod_Instance
    это первое апреля --- InstancePointer + MonthDay
    ??? --- InstancePointer + JustMentioned_Situation
    это письмо другу  --- InstancePointer + InstanceWithAddr
    """
    head_classes = [
        'Nominal',
        'PrepModNominal',
        'PersonByTitle',
        'PersonByTitle_WomanMasc',
        'GeoObjectByClass',
        'GeoObjectByClassGoverned',
        'GeoName',
        'OrgByType',
        'OrgByName',
```

```
        'AttrClauseModInstance',
        'AttrClauseModNominal',
        'NonSitReflInstanceOperCall',
        'SingularNumeralGroup',
        'QuotedInstance',
        'InstTitle',
        'IsolatedPart_Mod_Instance',
        'MonthDay',
        'JustMentioned_Situation',
        'InstanceWithAddr',
        'ProGenitival',
        'NumberPercent',
        'NumberPercentWithPrepositional',
        'InstanceColonWithInstanceGroup',
        'OneOfInstance',
        'ModalNominal',
        'InstanceWithDetachedAdjunct'
    ]
```

Классы для описания процессуальной именной группы с обстоятельством

```
class Process(SyNode):
    head_classes = [
        'ProcessualNoun_With_Object',
        'ProcessualNoun_With_ProObject',
        'ProcessualNoun_With_TransSubject',
        'ProcessualNoun_With_IntransSubject',
        'ProcessualNoun_With_Circ',
        'ProcessualNominal_With_Circ',
        'ProcAddr',
        'ProcInstr']

class ProcessualNominal_With_Circ(SyNode):
    """
    Сочетание процессуальной именной группы с обстоятельством
    Пример:
    шум дождя по крыше
    грохот дождя на железной крыше
    """
    head_classes = ['ProcessualNoun_With_IntransSubject']
    modifier_classes = ['Circumstance', 'PrepositionalCircumstance']
    needs_gram = {'isName': 0, 'anim': 0}
    specifier_order = Orders.Right()
    modifier_can_be_ellipsed = False

class ProcessualNoun_With_IntransSubject(SyNode):
    """
    Процессуальное существительное, соответствующее непереходному глаголу, с группой субъекта (процессуальное существительное с обстоятельством + группа субъекта процессуального существительного, соответствующего непереходному глаголу)
    Пример:
    взлет самолета
    """
```

```
        head_classes = ['ProcessualNounIntrans', 'ProcessualNoun_With_Circ',
'ProcessualNominal_With_Circ']
        argument_classes = ['ProcessualNoun_IntransSubject']
        specifier_order = Orders.Right()
```

Классы для описания деепричастного оборота

```
class ClauseWithGer(Method):
        head_classes = ['Clause']
        argument_classes = ['ClosedGerundial']
        specifier_order = Orders.Optional({
                "affirmative": Orders.Left(),
                "affirmative_emph": Orders.Right()
        })

class ClosedGerundial(Method):
        """
        Закрытое обособленное обстоятельство, выраженное деепричастным оборотом
(OpenGerundial + знак препинания)
        Пример:
        (засветил огонь и), накинув пальто, (выглянул наружу)
        (высовывались), рискуя свалиться в воду, (из отверстий разломанных крыш)
        """
        head_classes = ["OpenGerundial"]
        needs_gram = {'vf': 'ger'}
        argument_classes = [
                'CommaArg',
                'IsolatedPhraseEndArg',
                'SentenceEndArg',
                'TextEndArg']

class OpenGerundial(Method):
        """
        Открытое необособленное обстоятельство, выраженное деепричастным оборотом
        Пример:
        (засветил огонь и), накинув пальто (, выглянул наружу)
        (высовывались), рискуя свалиться в воду (, из отверстий разломанных крыш)
        """
        head_classes = ['Verbial', 'Verbial_TimeBefore', 'Verbial_TimeAfter']
        needs_gram = {'vf': 'ger'}
        argument_classes = [
                'CommaArg',
                'IsolatedPhraseStartArg',
                'SentenceStartArg']
```