ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

 «САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

на тему:

**Вычисление сходства русских текстов на основе синтаксических структур**

основная образовательная программа бакалавриата по направлению подготовки 45.03.02 «Лингвистика»

Исполнитель:

Обучающийся 4 курса

Образовательной программы «Прикладная,

экспериментальная и математическая

лингвистика (английский язык)»

Профиль «Прикладная, экспериментальная

и математическая лингвистика»

очной формы обучения

Щербинина Арина Александровна

Научный руководитель:

 к.ф.н., доц. Азарова И.В.

Рецензент:

к.ф.н., ст. преп. Добров А.В.

 Санкт-Петербург

 2018

**Содержание**

[Введение 4](#_Toc515907973)

[Глава 1. Линейно-структурная организация предложения 8](#_Toc515907974)

[1.1.Определение простого предложения 8](#_Toc515907975)

[1.2.Понятие структурной схемы 8](#_Toc515907976)

[1.3. Структурные схемы простых предложений 10](#_Toc515907977)

[1.4. Распространители простого предложения 12](#_Toc515907978)

[1.5. Коммуникативная организация предложения. Актуальное членение 16](#_Toc515907979)

[1.6. Топологические поля 17](#_Toc515907980)

[Глава 2. Семантико-синтаксический алгоритм сравнения предложений. Представление лексических значений слова в словаре RussNet 21](#_Toc515907981)

[2.1. Семантико-синтаксический алгоритм сравнения предложений для китайского языка 21](#_Toc515907982)

[2.2. Создание параллельного корпуса парафразов 23](#_Toc515907983)

[2.2. Организация лексических значений слова в словаре RussNet 26](#_Toc515907984)

[Глава 3. Разработка автоматической процедуры вычисления сходства текстов на основе синтаксических структур 30](#_Toc515907985)

[3.1. Материалы исследования. Тестовые выборки 30](#_Toc515907986)

[3.2. Первичная обработка текстов. Морфологический анализатор 30](#_Toc515907987)

[3.3. Чанкер и выделение топологических полей 32](#_Toc515907988)

[3.4. Алгоритм сравнения текстов 35](#_Toc515907989)

[3.5. Вычисление сходства новостных текстов. Оценка работы алгоритма 38](#_Toc515907990)

[Заключение 42](#_Toc515907991)

[Список литературы 45](#_Toc515907992)

[Электронные источники 49](#_Toc515907993)

[Приложение 50](#_Toc515907994)

[Код представленного в данной работе алгоритма 50](#_Toc515907995)

[Обозначения компонентов структурных схем 50](#_Toc515907996)

[Корпус новостных статей и параллельный корпус изложений 50](#_Toc515907997)

[Список парафразов, составленный на материале корпуса изложений 51](#_Toc515907998)

[Таблица 10. Оценка сходства текстов из новостного корпуса 54](#_Toc515907999)

# Введение

Автоматическое извлечение информации всегда было и до сих пор является важным направлением в научных работах по компьютерной обработке естественного языка. Многие исследователи собирают корпуса текстов для проверки тех или иных лингвистических гипотез, для машинного обучения и выполнения различных других задач. Среди всего многообразия корпусов отдельно выделяются одноязычные корпуса парафразов, интерес к которым в последнее время возрастает. Это можно объяснить широкой областью их практического применения: данные корпуса активно используются в программах по автоматическому реферированию [20] или упрощению текстов (например, для нужд таких проектов как Simple Wikipedia [47]), в автоматических вопросно-ответных системах, при решении задач машинного перевода, классификации различных текстов и т.д.

Для составления одноязычных параллельных корпусов необходима справедливая оценка сходства входящих в них текстов. Существует ряд алгоритмов по сравнению предложений/текстов на различных языках, однако среди них нельзя выделить какой-либо один, широко используемый и общепризнанный. Помимо уже названных областей сравнительные алгоритмы применяются для улучшения программ по выявлению плагиата – крайне актуальной для нашего времени задачи, – а также могут быть использованы различными средствами массовой информации для определения новизны новостей.

Предложения (а, следовательно, и тексты) можно сравнивать по трем аспектам: семантически; на уровне синтаксических структур и на уровне прагматики. Последняя область особенно сложна для изучения и практического воплощения и является целью всех разработчиков алгоритмов сравнения – пока недостижимой [25]. Что касается двух других аспектов, то изначально при сравнении учитывалась только лексическая семантика: тексты/предложения представлялись в виде «мешков слов» [20; 33]. Затем были предприняты попытки вычисления сходства предложений с опорой лишь на синтаксические структуры [31]. В ряде последних работ показана значимость и целесообразность объединения двух этих методов. Это способствует улучшению качества сравнения и получению более точных результатов, приближенных к человеческой оценке [25; 26].

Современные алгоритмы учитывают семантику, частеречные теги рассматриваемых слов, их порядок в предложении и проч. [10; 30] Семантическое сравнение можно проводить разными способами: лемматизировать словоформы или же брать их в том виде, в каком они встречаются в текстах/предложениях [28; 34]; учитывать только полное совпадение элементов [26] или прибавлять к нему различные семантические отношения (синонимия, гипонимия, антонимия и т.д.) [17; 26; 27; 29].

Что касается синтаксической составляющей, у прикладных лингвистов часто возникает вопрос, что понимать под синтаксической структурой и какую минимальную единицу выбрать на данном уровне сравнительного анализа. Разные исследователи решают эту проблему по-разному: одни считают сходство предложений в целом [31], другие разбивают предложения на более мелкие значимые непересекающиеся синтаксические единицы – чанки [Abney 1991] (в русской традиции эквивалентные синтагмам) – и вычисляют показатель сходства двух предложений путем сложения показателей сходства этих единиц [25; 26].

Если для языков с фиксированным порядком слов выделение чанков или более мелких синтаксических единиц не представляет особого труда, то языки со свободным порядком, такие как русский, немецкий и др., представляют определенные трудности, а значит, являют многообразие способов разбиения предложения на составляющие.

Таким образом, проблема проведения качественного семантико-синтаксического сравнительного анализа невероятно трудна и многогранна. Этим объясняется отсутствие общепринятого алгоритма сравнения не только для ряда различных, но и, зачастую, для одного конкретного языка.

Целью данной работы является разработка алгоритма сравнительного анализа текстов на русском языке, учитывающего семантическое сходство составляющих их слов и опирающего на синтаксические структуры.

Для достижения цели были поставлены следующие задачи:

* Представить теоретическую базу описания структуры простого предложения;
* Рассмотреть синтаксические единицы, выделяемые на высших уровнях синтаксического анализа в разных языках;
* Ознакомиться с существующими семантико-синтаксическими алгоритмами сравнения предложений для различных языков;
* Описать структуру представления лексических значений слова в компьютерном тезаурусе RussNet [48] и выделить семантические отношения для использования в практической части данной работы;
* Разработать семантико-синтаксический алгоритм сравнения текстов на русском языке:
	+ Разработать процедуру выбора однозначного морфологического анализа применительно к данной задаче;
	+ Разработать процедуру деления предложения на семантико-синтаксические фрагменты;
	+ Реализовать автоматическое выделение синтаксических связей между компонентами предложения;
	+ Задействовать при сравнении текстов информацию о семантических отношениях между леммами;
* Подобрать весовые коэффициенты для сопоставления синтаксических фрагментов разного типа методом регрессионного анализа, используя обучающую совокупность текстов изложений;
* Создать золотой стандарт оценок сходства для ряда новостных текстов;
* Протестировать работу программы на текстах золотого стандарта.

Материалом исследования выступает фрагмент НКРЯ [40] объемом 5 млн. словоупотреблений, состоящий из газетных статей, текстов законов и договоров а также научно-популярных и художественных текстов.

В качестве тестовой выборки для обучения и проверки результатов представленного алгоритма сравнения использовались школьные изложения (4 тыс. словоупотреблений) и новостные статьи (10 тыс. словоупотреблений).

Теоретическая значимость работы определяется разработкой и тестированием процедуры сравнения лексико-синтаксического компонента текстов. Практическая значимость исследования состоит в выработке метода выделения в текстах значимых семантико-синтаксических фрагментов, которые упрощают разбор линейных синтаксических структур, и подборе для них весовых коэффициентов, позволяющих получить справедливую оценку сходства при сопоставлении.

Данная работа состоит из введения, трех глав, заключения, списка литературы из 36 названий, перечня электронных источников из 13 пунктов и пяти приложений. В первой главе описываются теоретические вопросы, связанные с общей структурой построения простого предложения, как то структурные схемы и их распространители, а также представлена модель топологических полей для немецкого языка. Вторая глава посвящена рассмотрению существующего семантико-синтаксического алгоритма сравнения предложений на китайском языке, описанию эскперимента по составлению параллельного корпуса парафразов на материале изложений и организации лексических значений слова в компьютерном тезаурусе русского языка RussNet. В третьей главе представлен новый семантико-синтаксический алгоритм сравнения текстов на русском языке, рассмотрены результаты его применения к корпусу новостных текстов, а также дана оценка этих результатов.

# Глава 1. Линейно-структурная организация предложения

## 1.1.Определение простого предложения

Ряд существующих грамматик (традиционная, теньеровская [7], грамматика зависимостей и грамматика непосредственных составляющих) практически единогласно признает предложение максимальной синтаксической единицей. Однако, их взгляды на то, что следует выделять в качестве единицы минимальной, расходятся. Это объясняется сложностью организации предложения и функциями, выполняемыми им в речи.

В «Русской грамматике» [8] приводится следующее определение: «Простое предложение – это синтаксическая конструкция, которая является относительно независимым от внешнего и внутреннего контекста высказыванием, обладающим коммуникативной функцией и интонационным оформлением и включающим грамматически связанные формы слов. Оно построено на основе отвлечённого грамматического образца (структурной схемы), специально предназначенного языком для построения отдельных единиц. Кроме этого в предложении выражено грамматическое значение предикативности (соотнесённости ситуации с некоторым реальным или нереальным планом действительности, которая обнаруживается в системе противопоставленных рядов форм). Предложение обладает коммуникативной перспективой, которая выражается в линейной структуре при помощи разных средств: интонации, порядка слов, частиц и других модальных компонентов, а также имеет иерархически организованную семантическую структуру и типовой набор распространителей».

## 1.2.Понятие структурной схемы

В предложении в обязательном порядке реализуется субъектно-предикатная структура, которая является его семантической интерпретацией. Она, как и грамматическое значение предикативности, заключена в его ядре – структурной схеме. Существуют две принципиально различные трактовки этого понятия [Евсеева и др. 2007]:

* Н.Ю.Шведова понимает структурную схему как предикативный минимум: в нее не включаются компоненты, которые не образуют предикативную основу предложения, то есть его второстепенные члены. Структурные схемы в этой концепции записываются в виде неких формул, в которых определенными символами[[1]](#footnote-1) по морфологическим признакам (часть речи, ее грамматическая форма и т.д.) обозначены компоненты схем.
* Т.П.Ломтев, Г.А.Золотова и др. считают структурную схему номинативным минимумом. Такое понимание обращено не только к формальной устроенности предложения, но и к его смысловой организации. Согласно этой концепции, в состав структурной схемы входят не только главные члены предложения, но и обязательные второстепенные члены.

В данной работе структурные схемы считаются предикативным минимумом. В рамках этой концепции они разграничиваются по формальному устройству, семантике, парадигматическим свойствам построенных по этой схеме предложений, системе регулярных реализаций, правилам распространения.

Семантика структурной схемы предложения образуется взаимным действием следующих факторов: 1) грамматическими значениями ее компонентов в их отношением друг к другу (в однокомпонентных схемах — грамматическим значением компонента схемы); 2) специфическими для данной схемы лексико­семантическими характеристиками слов, в конкретных предложениях занимающих позиции ее компонентов.

В «Русской грамматике» принята классификация, в которой первичным основанием является деление структурных схем на свободные и фразеологизированные. Свободные схемы (их большинство, и они занимают центральное место в системе простого предложения) делятся на двукомпонентные и однокомпонентные. Двукомпонентные схемы, в свою очередь, подразделяются на схемы со спрягаемой формой глагола и без нее. Внутри схем со спрягаемой формой глагола выделяются схемы подлежащно­сказуемостные и не подлежащно­сказуемостные. Внутри класса схем без спрягаемой формы глагола вычленяются схемы с лексически не ограниченными компонентами — подлежащно­сказуемостные и не подлежащно­сказуемостные — и схемы с компонентами, ограниченными лексико­семантически.

Однокомпонентные схемы делятся на схемы со спрягаемой формой глагола и схемы без спрягаемой формы глагола (именные, инфинитивные и наречные). Фразеологизированные схемы классифицируются по грамматическому характеру лексически закрытого компонента: это фразеологизированные типы предложений с союзами, с предлогами, с частицами, с междометиями и с местоимениями.

В свободных двукомпонентных схемах словоформы находятся друг с другом в синтаксических отношениях. В большинстве случаев это одновременно и отношения между центральными семантическими компонентами предложения — субъектом и его предикативным признаком

## 1.3. Структурные схемы простых предложений

Всего в «Русской грамматике» приводится 30 структурных схем. Ниже подробно рассматриваются свободные структурные схемы, используемые в практической части данной работы (в двукомпонентных схемах особо выделен носитель предикативного признака).

Двукомпонентные структурные схемы:

* $ N\_{1} V\_{fin}$. Самая частотная структурная схема. В ней реализуется координация между подлежащим и сказуемым (сказуемое повторяет форму числа, иногда рода, подлежащего). Выражает отношение между субъектом и его предикативным признаком, обозначающим действие или состояние. Наличие объекта подразумевается глаголом: «*Лес шумит*», «*Дети веселятся*»;
* [$N\_{3}$]$ V\_{3s }$***Inf*.** Глагол указывает на модальную характеристику предложения, сам по себе он передаёт пропозицию. Обозначает модальную квалификацию (необходимость, желательность, предопределённость и т.д.) отвлечённо представленного действия или состояния. Эта модальная квалификация может быть субъектной (есть валентность на субъект [$N\_{3}$]): «*Мне хочется уехать*», «*Вам следует подождать*»;
* $N\_{2} \left[neg\right] V\_{3s}$. Выражает отношение между субъектом и его предикативным признаком – процессуальным состоянием: «*Воды убывает*», «*Несчастья не случилось*»;
* $N\_{1} Adj\_{1}$\*. В этой схеме реализуется координация между подлежащим и сказуемым по числу и роду. Обозначает отношение между субъектом и его предикативным признаком – непроцессуальным состоянием: «*Ребенок послушен*», «*Ночь тиха*»;
* $N\_{1} $***Part*\***. Выражает отношение между субъектом и предикативным признаком – состоянием, являющимся результатом действия: «*Дома восстановлены*», «*Обещания забыты*»;
* ***Praed*** *Inf.* Значение: наличие субъектно-модального состояния, осуществление действия или состояния. Есть валентность на субъект модального состояния – [$N\_{3}$]: «*Нам пора ехать*», «*Нельзя останавливаться*»;
* [*neg*] ***Praed*** $N\_{2}/N\_{4}$. Обозначает наличие субъектного состояния, обращенного к объекту. Есть валентность на субъект – [$N\_{3}$]: «*Видно следы*», «*Ему не видно следов*»;
* $Praed\_{part}$ $N\_{2}$. Обозначает отношение между субъектом состояния (он же – объект действия) и его предикативным признаком – состоянием как результатом действия: «*Наготовлено запасов*», «*Подтверждение не получено*»;

Однокомпонентные:

* $ V\_{3s}$. Значение: существование бессубъектного или субъектного состояния или процесса: «*Светает*», «*Знобит*»;
* $ V\_{3p}$. Обозначает существование действия или процессуального состояния, отнесенного к неопределенному субъекту: «*Стучат*», «*Кричат*»;
* $N\_{1}$. Значение: существование предмета или предметно представленного действия/состояния: «*Ночь*», «*Тишина*»;
* *Praed*! Обозначает существование субъектного/бессубъектного состояния, которое является результатом предшествующего действия: «*Закрыто!*», «*Окрашено*».

## 1.4. Распространители простого предложения

В предложениях, реализованных в речи, для увеличения их информативности помимо структурной схемы часто присутствуют различные распространители. «Распространение предложения — это введение в его состав таких членов, которые, не являясь компонентами структурной схемы предложения (предикативной основы), расширяют его состав» [8].

Распространители могут относиться как к предложению в целом, так и к какому-либо его компоненту. И в том, и в другом случае между распространителем и распространяемой единицей возникают подчинительные синтаксические связи [Белошапкова и др. 1999]. Они могут быть обязательными и необязательными. В первом случае из-за информативной недостаточности главного компонента между ним и зависимым устанавливаются комплетивные отношения.

Связь распространителей и тех компонентов, к которым они относятся, при передаче линейной структуры предложения можно отражать различными способами: например, в традициях грамматики непосредственных составляющих или грамматики зависимостей.

В зависимости от типа и характера этой связи разграничиваются распространяющие члены с объектным значением, с субъектным значением и распространяющие члены — определители.

Распространяющие члены с объектным и определительным значениями могут относиться как к главному члену предложения, так и к какому­либо другому распространяющему члену или к предложению в целом. Определители также могут относиться одновременно к двум членам предложения и выступать в качестве распространителей с двунаправленными связями. Распространяющие члены с субъектным значением чаще всего относятся к предложению в целом.

Распространяющие члены предложения имеют разнообразную формальную организацию: это могут быть отдельные формы слов, словосочетания — простые, сложные или комбинированные, – или ряды форм общих по своей синтаксической роли в предложении. Синтаксические связи и отношения, возникающие в распространенном предложении, имеют свои характеристики, которые во многих случаях не совпадают с характеристиками присловных подчинительных связей и отношений.

С опорой на [8] можно выделить следующие типы распространителей простого предложения:

* Присловный распространитель: *«У девочки были голубые глаза»;*
* Местоименный распространитель (прономинализация). Используется для внутритекстовой связи: «*У него не было иголки, и я дал ему свою с ниткой*»;
* Распространитель с двойным подчинением. Стоит в предикативно-атрибутивной позиции, как правило, выражен именем существительным или прилагательным: «*Он застал отца больным/больного*», «*Я знал его еще мальчиком*»;
* Свободный определяющий член. Может быть выражен:
	+ $N\_{5}$ в квазипассивных конструкциях, обозначает деятеля: «*Дом строится рабочими*»;
	+ $N\_{5}$ в безличных предложениях, обозначает стихийную силу: «*Волной опрокинуло лодку*»;
	+ $PP$ с $N\_{5}$ для обозначения совместности: «*Мне с друзьями не скучно*»;
	+ $PP$ с $N\_{4}/N\_{6}$, стандартное значение $N\_{4}$ – направление, $N\_{6}$ – местоположение; используется в том числе с отвлеченными существительными: «*Ребенку в темной комнате страшно*», «*Старик в своей растерянности жалок*»;
* Распространяющий инфинитив. Представляет собой упрощенный вариант сложноподчиненного целевого предложения: «*В кино ходить времени не хватает*»;
* Детерминант. Стоит в препозиции и выделяется паузой, может содержать в себе тему предложения или служить в качестве ситуатива, позволяющего понять тема-рематическое членение. Выражает субъектные, субъектно-обстоятельственные, объектные, объектно-обстоятельственные и обстоятельственные значения: «*У меня нет времени*», «*Дома за нее заступиться некому*», «*Сыну он хочет только добра*», «*В семью сообщили о моем приезде*», «*На столе лежит книга*»;
* Обращение (апеллятив). Чаще всего выражено именем, также может содержать обозначение лица с определенной квалификацией: «*Вась*», «*Подруга дней моих суровых, голубка дряхлая моя…*»;
* Вводная конструкция. Выражает отношение говорящего к высказыванию или степень его уверенности в своих словах: «*Все, вероятно, были в некотором замешательстве*»;
* Сравнительный оборот. По значению выделяют собственно сравнительные, репрезентирующие, градационные и приравнивающие: «*А кисти сочные как яхонты горят*», «*Хотел бы я иметь такого ученика как ты»*, «*Здесь было светлее, чем в поле»*, «*Грибники, как и рыболовы, не любят лишнего шума»*;
* Полупредикативные обороты. Содержат в себе пониженную в статусе предикацию, которая находится к таксисном отношении со сказуемым:
	+ Деепричастный. Чаще всего выражает обстоятельственное значение (времени, причины, уступительное, условное), также может означать сопутствующее действие или состояние: «*Поднявшись на лестницу, я увидел бухту*», «*Кошка дремала, поджав под себя передние лапы*»;
	+ Причастный. Выражает значение времени/причины, уступки или сопутствующего состояния: «*Согретый солнцем, мох зацветает*», «*Улицы, созданные пешеходами, перешли во власть автомобилистов*»;
	+ Субстантивный. Всегда относится к подлежащему; если стоит в препозиции, обозначает причину, уступку или сопутствующее качество, в постпозиции имеет характер пояснения, выражает дополнительное сообщение: «*Человек клана, он знает о кланах всё»*, «*Приезжал водовоз, старичок в кумачовой рубахе*»;
	+ Адъективный. Может относиться не только к подлежащему, но и к косвенному обозначению субъекта: «*Малообщительный по натуре,* ***он*** *стал еще нелюдимее*», «*Ивану Карловичу, усталому, совершенно не хочется ехать в клуб*»;
* Поясняющие обороты. Всегда стоят в постпозиции, относятся к второстепенным членам предложения:
	+ Субстантивный. Относится к существительному, местоимению (не являющимся субъектом предложения), прилагательному, наречию: «*Я остановился у приятеля отца, начальника почтовой конторы*», «*Он был в заграничных, в цвет костюма, туфлях*», «*Он никогда, даже в молодости, не производил впечатление здорового*»;
	+ Адъективный. Появляется в рядах, когда передается одна и та же характеристика: «*Светлые, почти белые волосы прилипли ко лбу*»;
	+ Наречный. Может относиться к словосочетанию или наречию: «*После конференции, вечером, Федор созвал совещание*», «*Мы к этому успеху относились свысока, снисходительно*»;
	+ Со спрягаемой формой глагола. Одно сказуемое как бы поясняет другое: «*Надо писать о том, что «горит», не ждет»*;
	+ Инфинитивный: *«У профессора была заветная мечта: установить возраст китов»,* «*Многие пришли просто так: посидеть да побеседовать*».

## 1.5. Коммуникативная организация предложения. Актуальное членение

В предложении необходимо различать конструктивный (формальный) и коммуникативный аспекты [Белошапкова и др. 1999]. Для выяснения формальной организации нет необходимости анализировать и учитывать что-либо, кроме самого предложения; а коммуникативный аспект, напротив, проявляется в контексте речевой ситуации и тесно связан с актуальным членением. В языках с фиксированным порядком слов формальная структура не зависит от тема-рематической организации предложения: тему обычно выражает подлежащее, рему – сказуемое, в иных случаях рема, как правило, выделяется при помощи фразового ударения. В языках со свободным порядком слов больше способов выделить новую информацию: так, говорящий имеет возможность переставить слова в зависимости от своих коммуникативных целей (ср. «*Брат приехал*» – «*Приехал брат*»).

Все это приводит к невозможности однозначного отображения структуры предложения с помощью непосредственных составляющих: для каждого предложения в речи существует множество вариантов формальной организации, число которых возрастает пропорционально количеству слов. Кроме того, языки со свободным порядком слов представляют определенные трудности и для анализа контекстно свободной грамматики: в предложении могут встретиться «разорванные» единицы, которые невозможно отразить с помощью деревьев словосочетаний, избегая пересечения ветвей.

## 1.6. Топологические поля

Утверждение о том, что в «свободных» языках отсутствуют какие-либо ограничения на порядок слов неверно, и в большинстве случаев можно выделить следующую закономерность построения предложения: сначала идет некое поле, содержащее в себе тему, за ним следуют различные клитики, предикат и прочие распространители (а также придаточные предложения – при наличии) [15]. Данное наблюдение позволяет при высокой вариативности размещения элементарных составляющих выделить в предложении более крупные единицы, занимающие одну и ту же позицию относительно его предикативного центра.Такие единицы получили название топологических полей. Этот термин был введен [Hoehle 1983] для немецкого языка. В его работе топологическим называется линейное поле, выделяемое на высших уровнях синтаксического анализа и являющееся частью большего синтаксического единства (например, клаузы). Топологическое поле не обязано быть семантически связным, а на число слов, входящих в него, можно наложить ограничение.

В более поздней работе [Hoehle 1986] представлена модель топологических полей для немецкого и некоторых других германских языков. Согласно ей, простые предложения (в том числе в составе сложного) делятся на три типа: verb-last (VL), verb-first (V1) и verb-second (V2). К VL относятся придаточные предложения; V1 включает в себя императивы, вопросы, на которые можно ответить да/нет, и придаточные условные; V2 состоит из утвердительных предложений и придаточных, вводимых союзными словами. Для каждого типа предложений предусмотрена своя модель расположения топологических полей (см. Таблицу 1).

Таблица 1. Порядок расположения топологических полей для немецкого языка

|  |  |
| --- | --- |
| Тип предложения | Топологические поля |
| VL | (KOORD) (C) (MF) VC (NF) |
| V1 | (KOORD) (LV) LK (MF) (VC) (NF) |
| V2 | (KOORD) (LV) VF LK (MF) (VC) (NF) |

* VF (Vorfeld или «пре-поле») – первый элемент в предложениях типа V2. Часто является темой предложения, однако не обязательно выполняет только эту функцию (например, может быть и ремой);
* LK (Linke Klammer или «левая граница») – поле для личных глаголов в предложениях типа V1 и V2. В предложениях типа VL заменяется на поле C, включающее в себя предикат придаточного предложения;
* MF (Mittelfeld или «среднее поле») – необязательное поле, ограниченное слева полем LK, а справа глагольным комплексом VC или NF. Сюда попадает большинство глагольных дополнений, наречий и предложных конструкций (в том случае, если они не были вынесены в интерпозицию – поле VF – или не являются просодически тяжелыми, т.е. не относятся к полю NF);
* VC – поле глагольного комплекса. Включает в себя личные и неличные глаголы из предложений типа VL;
* NF (Nachfeld или «пост-поле») содержит просодически тяжелые элементы, такие как вынесенные в пост-позицию предложные конструкции или придаточные определительные предложения;
* KOORD (Koordinationsfeld или «координирующее поле») – поле для союзов, стоящих между частями сложного предложения;
* LV (Linksversetzung или «левое поле») включает в себя все конструкции, расположенные слева от предиката [Frey 2004].

Модель топологических полей прежде всего дистрибутивная. В ней не отражается предикативная структура предложения и не представлены отношения между единицами внутри полей. Несмотря на это, использование данной модели на начальном этапе анализа значительно улучшает его точность [Meurers 2002]. Топологические поля представляют скелет предложения и таким образом уменьшают неоднозначность. Без их предварительного выделения неясно, где проходят границы клауз в сложных предложениях, и к какому предикату относятся те или иные распространители.

Очевидно, что данную модель с некоторыми корректировками можно применить и для русского языка. Так, например, в поле VF попадут детерминат или подлежащее, носители предикативного признака предложения будут относиться к одному из полей LK или VC, а свободный определяющий член – к MF или LV, подчинительные союзы явно будут отнесены к полю KOORD, присловные распространители, являющиеся прямыми дополнениями глагола-сказуемого, и распространяющий инфинитив – к полю VC, а полупредикативные обороты – к полю NF.

На основании этого мы предлагаем выделять следующие типы полей:

1. Поле сказуемого;
2. Поле подлежащего;
3. Поле детерминанта, куда входят все распространители, стоящие слева от подлежащего при прямом порядке слов;
4. Поле свободного определяющего члена;
5. Поле распространителей, относящихся к сказуемому и в предложении стоящих справа от него;
6. Поле подчинительных союзов.

Тот факт, что все перечисленные выше распространители могут быть выражены словосочетаниями, позволяет внутри обозначенных топологических полей выделить более мелкие единицы – синтагмы или чанки, – составляющие которых связаны друг с другом подчинительными синтаксическими связями.

# Глава 2. Семантико-синтаксический алгоритм сравнения предложений. Представление лексических значений слова в словаре RussNet

## 2.1. Семантико-синтаксический алгоритм сравнения предложений для китайского языка

Синтаксические структуры играют важную роль в определении сходства предложений на естественном языке. Несмотря на то, что их использование данной области в отрыве от семантики не имеет особого практического смысла и дает трудно-, а порой и вовсе неинтерпретируемые результаты, учет лексического содержания вместе с его структурной организацией значительно улучшает качество сравнения. Комбинировать семантический и синтаксический компоненты можно различными способами, ниже приведен один из них – алгоритм сравнения предложений, разработанный [Li, Li 2015] для китайского языка.

 Для вычисления семантического сходства слов, составляющих предложения, [Li, Li 2015] используют платформу HowNet [Dong, Dong 2014]. В ней слова китайского языка представлены одним или несколькими понятиями, которые, в свою очередь, описаны с помощью набора семем (элементарных значимых единиц, составляющих понятие, у каждой из которых своя роль). Для многозначных слов, представленных несколькими понятиями, считается сходство всех их значений, а затем выбирается максимальный показатель, который отражает сходство между этими словами:

 $Sim\left(W\_{1}, W\_{2}\right)= \max\_{i=1…m,j=1…n}Sim(S\_{1i}, S\_{2j})$, где $W\_{1},W\_{2}$ – слова, m и n – число значений первого и второго слова соответственно, $S\_{11},S\_{12},…, S\_{1m}$ – значения первого слова, $S\_{21},S\_{22},…, S\_{2n}$ – значения второго слова.

Очевидно, что вычисление сходства двух понятий/значений может быть сведено к вычислению сходства составляющих их семем. В HowNet оно рассчитывается как семантическое расстояние по формуле

 $Sim\left(p\_{1},p\_{2}\right)=\frac{a}{d+a}$, где $p\_{1},p\_{2}$ – семемы, *a* – регулируемый параметр, *d* – длина пути из $p\_{1}$ в $p\_{2} $в иерархическом дереве семем.

В HowNet представлены 8 типов отношений между семемами, однако при вычислении их сходства в стандартном алгоритме учитывается только отношение гипонимии, и значения таким образом варьируются в диапазоне [0, 1]. [Li, Li, 2015] учли также отношение антонимии, тем самым расширив диапазон до [-1, 1]: для антонимов используется формула $Sim\left(W\_{1}, W\_{2}\right)=-Sim(S\_{1},S\_{2})$.

Синтаксический анализ упрощает автоматическую процедуру вычисления сходства предложений и способствует получению более точных результатов. Сходство целого складывается из сходства составляющих, поэтому вместо того, чтобы сравнивать предложения целиком, с помощью метода центрального компонентного анализа (инструмент LTP [24] для китайского языка) их делят на три части: группу подлежащего, группу сказуемого и группу дополнений, – представляющие собой так называемый скелет предложения.

Таким образом общее сходство предложений вычисляется по следующей формуле:

$$Sim\left(S\_{1},S\_{2}\right)=β\_{1}∙β\_{2}$$

$$β\_{1}=k⋅λ⋅γ⋅φ$$

$$β\_{2}=[α\_{1}Sim\_{1}\left(B\_{1},B\_{2}\right)+α\_{2}\left|Sim\_{2}\left(B\_{1},B\_{2}\right)\right|+α\_{3}Sim\_{3}\left(B\_{1},B\_{2}\right)]$$

Значение $Sim\left(S\_{1},S\_{2}\right)$ состоит из двух частей: $β\_{1}$ – корректирующий коэффициент, $β\_{2}$ – показатель семантического сходства. Значение $β\_{1}$ складывается из четырех параметров:

1. Параметр *k* нормализует значение сходства для предложений различных типов. Например, вопросительные предложения значительно отличаются по тону от остальных, поэтому при сравнении вопросительного предложения с императивом, простым повествовательным или восклицательным предложением *k* приравнивается к 0.1, при сравнении иных типов между собой k=0.5, и k=1 для предложений одного типа.
2. Параметр $λ$ учитывает количество элементов в сравниваемых предложениях и играет значительную роль, когда оно различно. Его значение вычисляется по формуле $λ=\frac{2∙i}{(m+n)}$, где *m* и *n* – число элементов в $S\_{1}$ и $S\_{2}$ соответственно, *i* – число элементов этих предложений, совпавших при сравнении.
3. 𝛾 – отрицательный параметр, он приравнивается к -1, если предикативные центры $S\_{1}$ и $S\_{2}$ связаны отношением антонимии или при полном сходстве предикатов в одном из предложений перед ним стоит отрицательная частица «не».
4. 𝜑 – дополнительный коэффициент, он приравнивается к -1, если предикаты двух предложений представлены антонимами, а также подлежащее и прямое дополнение в них меняются местами.

Так как после синтаксического анализа предложение разделено на три части, значение $β\_{2}$ складывается из $Sim\_{1}\left(B\_{1},B\_{2}\right)$, $Sim\_{2}\left(B\_{1},B\_{2}\right)$ и $Sim\_{3}\left(B\_{1},B\_{2}\right)$ – сходства подлежащих, сказуемых и распространителей соответственно. Значение $Sim\_{2}\left(B\_{1},B\_{2}\right)$ может быть отрицательным, поэтому оно берется по модулю. Параметры $α\_{1}=0.3$, $α\_{2}=0.5$ и $α\_{3}=0.2$ отражают значимость той или иной части предложения при сравнении.

В практической части нашего исследования данный алгоритм был взят за основу. Его модификация подробно описана в Главе 3. Принимая во внимание явные отличия русского языка от китайского мы считаем нужным подробно остановиться на семантических отношениях, связывающих лексемы русского языка, и их представлении в RussNet – аналоге китайской платформы HowNet.

## 2.2. Создание параллельного корпуса парафразов

В русском языке, в отличие от китайского, широко используются парафразы на уровне как лексики, так и синтаксических структур. Одну и ту же коммуникативную цель можно реализовать различными способами. Для определения наиболее типичных парафразов нами был проведен эксперимент по составлению параллельного корпуса на материале изложений отрывка из статьи В.М.Пескова «Родина Ломоносова», написанных учениками 8 класса общеобразовательной школы.

В корпус вошло 24 текста, включая оригинал статьи (суммарный объем 4 тыс. словоупотреблений). Выравнивание осуществлялось на уровне чанков и на уровне предложений. Фрагмент полученного корпуса приведен на Рисунке 1: зеленым цветом выделены синтаксические и лексические парафразы, красным – добавление информации, не присутствующей в оригинальном тексте, синим – перенос информации из одной части текста в другую. Ссылка на полный корпус и перечень выделенных синонимичных структур находятся в Приложении к данной работе.

 

Рисунок 1. Фрагмент параллельного корпуса изложений

Кроме того в рамках данного эксперимента нами было оценено сходство текстов изложений с оригиналом и между собой. Оценки выставлялись вручную на уровне предложений, при этом основное внимание уделялось семантике, однако синтаксичесое строение также учитывалось:

10 – 100%-е совпадение или же незначительные изменения синтаксической структуры (например, опущение/вставка подлежащего, присутствующего/отсутствующего в тексте оригинала, замена какого-либо слова на синонимичное сочетание и т.д.);

9 – более значительное изменение синтаксической структуры предложения (например, объединение его с последующим предложением) при полном сохранении его смысла;

8 – значительное изменение структуры предложения или добавление дополнительной информации, не содержащейся в оригинальном тексте, при сохранении исходного смысла (возможны незначительные опущения);

7 – информация данного предложения оригинального текста содержится в другом предложении изложения (не совпадает деление на предложения);

6 – присутствует лишь часть информации, содержащейся в тексте оригинала;

5 – часть информации сохранена, но находится в другой части текста изложения;

4 – смысл предложения передан не совсем правильно или в нем присутствует дополнительная информация при отсутствии исходного смысла;

3 – в предложении содержится дополнительная информация и его смысл передан неправильно;

1 – где-то в тексте изложения содержится совсем незначительная часть информации из данного предложения оригинального текста;

0 – информация не присутствует в изложении.

Распределение оценок показано в Таблице 2. Зеленым и синим выделены предложения, получившие максимальные оценки (10 и 9 соответственно), красным – минимальные (0). Для определения сходства текстов друг с другом была подсчитана корреляция проставленных оценок. Результаты представлены в Таблице 3, цветом выделены максимальные значения (большие или равные 0,6). Показатели корреляции использовались в практической части данного исследования, их применение подробно описано в Главе 3.

Таблица 2. Оценки соответствия изложений тексту-оригиналу



Таблица 3. Матрица корреляции изложений



## 2.2. Организация лексических значений слова в словаре RussNet

RussNet – это компьютерный тезаурус русского языка WordNet-типа. Он начал создаваться в 1999 году в качестве инициативного проекта кафедры математической лингвистики СПбГУ [Azarova 2008]. RussNet является оригинальной системой лексикализованных понятий русского языка, а не переводом тезауруса английского языка WordNet, и состоит из иерархических структур, образованных значениями четырех основных частей речи: глаголов, существительных, прилагательных и наречий. Единицами описания являются синсеты, в состав которых входят лексемы и устойчивые словосочетания, выражающие лексические значения, оформленные в русском языке. Между синсетами устанавливаются семантические отношения (гипонимические, меронимические, каузативные, пресуппозиционные), между лексемами — семантико-деривационные [Азарова 2015].

Элементы одного синсета связаны между собой отношением синонимии, при этом если они могут без потерь заменять друг друга в одном и том же контексте, например, *любой–каждый–всякий*, отношение обозначается как Synonym, иначе – Near\_Synonym: *задаваться–задавать*.

Между синсетами устанавливаются симметричные отношения антонимии (Atonym): *тепло–холодно* – и конверсии (Converse terms): *превосходить–уступать*, – а также несимметричные отношения гипо- и гиперонимии (Hyponym, Hypernym): *клён–дерево* – и меронимии с холонимией (Meronym, Holonym): *дерево–ствол*. Кроме того выделяются отношения лексического вывода: тропонимия (Troponym) – отношение между частным действием и более сложным: *купить–платить*, каузация (Cause): *убить–умереть*, пресуппозиция (Presupposition) – обозначение необходимой для данного действия предпосылки: *промахнуться–целиться*, *лязг–удар*, действие-часть (Subevent) – сходное с меронимией отношение при обозначении признаков: *храпеть–спать*.

В RussNet также включены отношения, обозначающие фазы ситуаций: инхоативы (Inchoative) – переход в состояние или начало действия: *сесть–сидеть*, *взлететь–лететь* – и терминативы (Terminative) – выход из состояния или завершение действия: *спать–проснуться*.

Видовые отличия глаголов обозначаются семантико-грамматическими отношениями Perfective\_aspect и Imperfective\_aspect: *приблизиться–приближаться, сдвинуться–сдвигаться.*

Для связи значений разных частей речи в концепции WordNet словарей используется префикс XPOS. Обычно таким образом передается значение транспозиции – выражение значения, характерного для одной части речи, при помощи другой: *любить–любовь*, *дерево–деревянный,* *шуметь–шум–шумный* и т.д. Поскольку при этом нарушается принятый тест на взаимную замену слов в контексте, отношение называется XPOS\_Near\_Synonymy.

Русский язык является высоко деривационным: словообразовательные связи задают множество семантически маркированных отношений. В RussNet для их обозначения используется специальный префикс DER с уточнением характера связи:

* DER\_Synonym – отношение, связывающее нейтральное слово и его экспрессивные дериваты: *старик–старичок*;
* DER\_Antonym – отношение, присоединяющее к маркированному значению, обозначающему положительный признак, комплементарный антоним: образования с приставками *не*-, *анти-*, *противо*-, *дез*-, *а*- и т.д.;
* DER\_Hyponym – отношение, в котором деривационный аффикс придает производному слову дополнительный смысловой оттенок, не сводимый к эмоционально-экспрессивному или стилистическому: *белый–беловатый*, *рука–ручища*;
* DER\_Role – образование существительных, обозначающих участников действия, через отсылку к глаголу: *сеять–сеятель*;
* DER\_Involved – образование глагольных значений через обозначение участника при помощи существительного: *рассупонить* – “Развязать *супонь* у хомута при распряжке лошадей”.

Последние два отношения могут быть уточнены в плане характеристики роли актанта: Agent, Subject, Instrument, Location, или Object.

В Таблице 4 приведен полный список отношений, представленных в RussNet.

Таблица 4. Семантические и семантико-деривационные отношения, представленные в словаре RussNet

| Название отношения | Определено на | Обратное отношение |
| --- | --- | --- |
| Симметричные отношения |
| Antonym | синсет<–>синсет | Antonym |
| Converse terms | синсет<–>синсет | Converse terms |
| Несимметричные отношения |
| Hypernym | синсет–>синсет | Hyponym |
| Holonym | синсет–>синсет | Meronym |
| Отношения лексического вывода |
| Troponym | синсет–>синсет | Trop\_hypernym |
| Cause | синсет–>синсет | Is\_caused\_by |
| Presupposition | синсет–>синсет | Is\_presupposed |
| Subevent | синсет–>синсет | Complex\_event |
| Инхоативные отношения |
| Inchoative | синсет–>синсет или слово-значение–>слово-значение | Initiated\_by |
| Terminative | синсет–>синсет или слово-значение–>слово-значение | Action/State |
| Словообразовательные отношения |
| XPOS\_Near\_Synonym | слово-значение–>слово-значение | Related\_to |
| DER\_Antonym | слово-значение–> слово-значение | Related\_to |
| DER\_Hyponym | слово-значение–>слово-значение | Related\_to |
| DER\_Troponym | слово-значение–> слово-значение | Related\_to |
| DER\_Role\_Agent | слово-значение–> слово-значение | Related\_to |
| DER\_Role\_Subject | слово-значение–> слово-значение | Related\_to |
| DER\_Role\_Object | слово-значение–> слово-значение | Related\_to |
| DER\_Role\_Instrument | слово-значение–> слово-значение | Related\_to |
| DER\_Role\_Location | слово-значение–> слово-значение | Related\_to |
| DER\_Involved\_Agent | слово-значение–> слово-значение | Related\_to |
| DER\_Involved\_Subject | слово-значение–>слово-значение | Related\_to |
| DER\_Involved\_Object | слово-значение–> слово-значение | Related\_to |
| DER\_Involved\_Instrument | слово-значение–> слово-значение | Related\_to |
| Семантико-грамматические отношения |
| Perfective\_aspect | слово-значение–> слово-значение | Imperfective\_aspect |
| Imperfective\_aspect | слово-значение–> слово-значение | Perfective\_aspect |

# Глава 3. Разработка автоматической процедуры вычисления сходства текстов на основе синтаксических структур[[2]](#footnote-2)

## 3.1. Материалы исследования. Тестовые выборки

Помимо корпуса изложений, описанного выше, в практической части данной работы использовались два корпуса текстов:

* Корпус новостных текстов, в который вошли 29 статей суммарным объемом в 10 тыс. словоупотреблений с порталов Lenta.ru [44], Gazeta.ru [43], РИА Новости [42] и сайта радио «Свобода» [41] на следующие темы:
	+ выборы президента Российской Федерации,
	+ отравление С.В.Скрипаля в Солсбери,
	+ высылка российских дипломатов из Великобритании;
* Фрагмент НКРЯ [40] объемом 5 млн. словоупотреблений, размеченный морфологически со снятой неоднозначностью. Тексты, каждый из которых не превышает объема в 5 тыс. словоупотреблений, представлены в следующем процентном соотношении: 60% — газеты, 15% — научно-популярные тексты, 15% — фрагменты художественных текстов, 10% — тексты различных документов (законы, договоры).

## 3.2. Первичная обработка текстов. Морфологический анализатор

Тексты обоих корпусов были подвергнуты морфологическому анализу с помощью инструмента Pymorphy2 [Korobov 2015]. Однако, его выдачу пришлось корректировать, потому как она не совсем подходила для выполнения задач, поставленных в данной работе.

Прежде всего, размер грамматического тега в Pymorphy2 зависит от части речи: для существительных отражается 5 характеристик (часть речи, одушевленность/неодушевленность, род, число, падеж), для глаголов – 6 (часть речи, вид, транзитивность, число, время, наклонение) и т.д. Кроме того, каждая характеристика обозначается четырьмя символами: в случае обработки и последующего хранения большого количества данных это значительно увеличивает объем необходимой памяти. В связи с этим совместно с К.В.Сипуниным c опорой на результаты диссертационного исследования А.Ю.Новиковой [Новикова 2017] была произведена унификация тегсетов и удаление из них избыточной для данной работы информации, для чего был написан скрипт на Python [44]. Теги всех частей речи приводились к единому формату, включающему в себя 7 позиций: 1) часть речи, 2) одушевлённость, 3) падеж, 4) число, 5) род, 6) лицо и 7) вид. Вместо отсутствующих либо нерелевантных (одушевлённость имен не в винительном падеже и проч.) граммем ставился прочерк, но при этом шесть прочерков подряд не ставилось: в тегсете для слов неизменяемых частей речи позиция всего одна. Для обозначения грамем использовались двусимвольные сокращения, соответствующие нотации А.Ю.Новиковой[[3]](#footnote-3).

Pymorphy2 выдает все варианты грамматического разбора слова, ранжируя их в соответствии с их условной вероятностью. Она вычисляется по материалам OpenCorpora [45] с использованием сглаживания Лапласа. Однако, такая оценка есть только для 20 тыс. слов, в иных случаях вероятность различных разборов либо считается равномерной (для словарных слов), либо оценивается на основе эмпирических правил (для несловарных слов). Таким образом, первый предложенный разбор верен примерно в 79% случаев [документация Pymorphy2].

Для выбора одного грамматического тега из полученного списка и снятия неоднозначности морфологической разметки по методу Э.Брилла [12; 13; 14] была написана программа на Python. Она учитывает локальный контекст и работает по триграммам: для выбора одного тега рассматриваемого слова принимаются во внимание теги предыдущего и последующего слов. Полученные триграммы сравниваются с рядом «золотых» триграмм, извлеченных из фрагмента Национального корпуса русского языка и ранжированных по частоте встречаемости в текстах, на основании чего отбрасываются все тегсеты рассматриваемого слова, кроме наиболее подходящего в данном контексте.

Кроме того, при разметке текстов для корректного выделения триграмм было принято решение сохранять знаки препинания. Они заносились в специальный тег <pc>, в атрибутах указывалось, является ли данный знак терминальным, нетерминальным или неоднозначным. К терминальным относились вопросительный и восклицательный знаки, точка с запятой, обычные и квадратные скобки; к нетерминальным – кавычки. Неоднозначными знаками препинания считались точка, многоточие, запятая, двоеточие и тире. В некоторых случаях неоднозначность снималась с привлечением локального контекста: так, точка относилась к терминалам, если стояла в абсолютном конце предложения, или к нетерминалам, если входила в состав распространенного общепринятого сокращения[[4]](#footnote-4); прочие знаки считались терминальными в позиции слева от союзов и союзных слов, способных разграничивать лишь простые предложения в составе сложного (не словосочетания)[[5]](#footnote-5). Таким образом, триграммы, центральным элементом которых являлся терминальный знак препинания, не рассматривались и не включались в «золотой» список; в этих случаях, как и в абсолютном начале предложения, для выбора грамматического тега использовались биграммы.

Результатом первичной обработки текстов являются xml-файлы, в которых сохраняется исходное деление на абзацы и знаки препинания с обозначенными выше характеристиками, а для каждого слова указываются лемма и грамматический тегсет.

## 3.3. Чанкер и выделение топологических полей

Из полученных файлов выделяются чанки – непересекающиеся значимые синтаксические единицы. Выделение происходит в два этапа: сначала каждый абзац разбивается на предложения или эквивалентные им единицы по терминальным знакам препинания, затем внутри этих структур находятся по грамматическим тегам чанки из «золотого» списка, извлеченного из корпуса изложений. Помимо грамматической информации в атрибуты выделяемых чанков записываются леммы входящих в него слов; также в каждом чанке определяется главное слово, представленное леммой и соответствующим ей тегом. В случае если чанк состоит из одного слова, это слово и считается главным.

На следующем этапе чанки объединяются в топологические поля. Предложенные нами в первой главе типы полей обозначались следующим образом:

* Coordination – поле для подчинительных союзов, соединяющих клаузы в сложноподчиненном предложении;
* LeftNounRem – распространители, стоящие слева от подлежащего при прямом порядке слов – детерминанты – или слева от сказуемого при инверсии;
* NomNoun – подлежащее с присловными распространителями;
* FreeDet – свободный определяющий член при прямом порядке слов или распространители, стоящие между сказуемым и подлежащим при инверсии;
* FinVerb – сказуемое с присловными распространителями;
* RightRem – поле для распространителей, стоящих справа от сказуемого при прямом порядке слов (справа от подлежащего при инверсии);
* UnableToDefine – поле, куда попадает предложение (или эквивалентная ему единица), если в нем нельзя выделить ничего из вышеперечисленного. Как правило, возникает при обработке предложений, построенных по редким (или фразеологизированным) структурным схемам.

Отнесение чанков к тому или иному полю происходит в два этапа. На первом этапе рассматриваются главные слова чанков и выделяются поля подчинительных союзов (главное слово – подчинительный союз/союзное слово), подлежащих (существительное или местоимение в именительном падеже) и сказуемых (личная форма глагола, краткое прилагательное, краткое причастие, пердикатив, компаратив). Затем, опираясь на их расположение в предложении, выделяются остальные поля (при наличии).

После выделения топологических полей анализируемый текст представляется как список предикатов: реальных и мнимых. Центральным элементом считается поле FinVerb, мнимые предикаты (с пустым полем сказуемого) выделяются в случае эллипсиса, назывных предложений, использования некоторых редких структурных схем или заполнения валентности подлежащего рядом – в этом случае реальный предикат относится к ближайшему к нему подлежащему, другим приписываются мнимые. В атрибуты каждого предиката заносятся: информация о соответствующем ему подлежащем (поле NomNoun или «None» при его отсутствии); матрица, состоящая из одной строки и семи столбцов, в которой при помощи нолей и единиц отражено наличие или отсутствие того или иного поля при данном предикате; сам массив этих полей и номер предложения, к которому данный предикат относится. Пример оформления текста и отдельного предиката приведен на Рисунке 2.



Рисунок 2. Представление текста в виде списка предикатов. Пример конкретного предиката

## 3.4. Алгоритм сравнения текстов

Тексты сравниваются по этим предикатам. По ходу сравнения заполняются числовыми значениями 6 матриц, строками в которых являются предикаты из первого текста, а столбцами – предикаты из второго. Первые две матрицы связаны с полями предикатов, третья и четвертая – с полями подлежащих, пятая и шестая – с полями распространителей. Матрицы 1, 3 и 5 состоят из нолей и единиц: 1 означает полное совпадение предикатов/подлежащих/распространителей соответственно, 0 – несовпадение или частичное совпадение. В случае частичного совпадения соответствующее ему числовое значение заносится в матрицу 2, 4 или 6 в зависимости от рассматриваемых полей; при полном совпадении или несовпадении полей в этих матрицах стоят ноли.

Общий принцип получения числовых показателей заключается в следующем. Чанки рассматриваемых полей последовательно сравниваются друг с другом: как правило, поля сказуемого и подлежащего состоят из одного чанка, количество чанков в полях распространителей может быть любым. Сначала сравниваются главные слова этих чанков[[6]](#footnote-6): если их грамматические тегсеты совпадают и леммы идентичны или являются полными синонимами, к показателю сходства[[7]](#footnote-7) прибавляется 1. В случае если тегсеты различны, а леммы одинаковы или связаны отношением синонимии (полной или частичной: отношения Synonym, Near\_Synonym, XPOS\_Near\_Synonym из RussNet), или являются видовой парой (Perfective\_aspect и Imperfective\_aspect), прибавляется 0,7. Затем по этому же принципу зависимые слова сопоставляются с главными и друг с другом, однако значения показателей берутся меньшие: 0,5 и 0,3 соответственно[[8]](#footnote-8). Стоит отметить, что, по аналогии с [Li, Li 2015], была предпринята попытка учитывать при сравнении отношение антонимии, однако, вопреки нашим ожиданиям, это способствовало ухудшению качества работы алгоритма.

Перед занесением в матрицу полученные параметры нормализуются. Для полей предикатов и подлежащих используется формула $s= \frac{2∙p}{l\_{1}+l\_{2}-i}$, где p – показатель сходства, $l\_{1}$ и $l\_{2}$ – длины полей, *i* – число элементов первого и второго поля, которые при сравнении не учитывались. Показатель сходства полей распространителей для каждого предиката получается по формуле

$$s=\frac{2∙\sum\_{\begin{array}{c}y=0,…,l\_{1},\\z=0,…,l\_{2}\end{array}}^{}\sum\_{\begin{array}{c}n=0,…,l\_{f1},\\m=0,…,l\_{f2}\end{array}}^{}\left(\frac{p\_{d}}{\left(l\_{ch1\_{n}}+l\_{ch2\_{m}}-i-2\right)\left(l\_{f1\_{y}}+l\_{f2\_{z}}\right)}+\frac{p\_{m}}{\left(l\_{f1\_{y}}+l\_{f2\_{z}}\right)}\right)}{\left(l\_{1}+l\_{2}-k\right)}, $$

где $p\_{d}$ – показатель сходства зависимых слов; $p\_{m}$ – показатель сходства главных слов; $l\_{1}$ и $l\_{2}$ – общее число полей при рассматриваемых предикатах (1 и 2 соответственно); *k* – целочисленный суммарный показатель, отражающий наличие при предикатах полей подлежащего и сказуемого, его значение варьируется в диапазоне [0, 4]: 0 – у обоих предикатов эти поля отсутствуют, 4 – присутствуют; $l\_{f1}$ и $l\_{f2}$ – число чанков в отдельно взятом поле; $l\_{ch1}$ и $l\_{ch2}$ – число слов в отдельно взятом чанке; *i* – число элементов из рассматриваемых чанков, которые при сравнении не учитывались.

После заполнения матриц элементы каждой из них суммируются между собой. Таким образом получается 6 параметров: $x\_{1}$ – число полностью совпавших предикатов, $x\_{2}$ – сумма показателей сходства частично совпавших предикатов, $x\_{3}$ – число полностью совпавших подлежащих, $x\_{4}$ – сумма показателей частично совпавших подлежащих, $x\_{5}$ – число полностью совпавших распространителей (в полях LeftNounRem, FreeDet, RightRem, UnableToDefine), $x\_{6}$ – сумма показателей частично совпавших распространителей.



Рисунок 3. Коэффициенты значимости параметров, из которых складывается сходство текстов

В целях нормализации, что особенно важно в тех случаях, когда размеры сравниваемых текстов значительно отличаются друг от друга, для каждого параметра используется формула $\frac{2∙x}{l\_{t1}+l\_{t2}},$ где *x* – параметр, $l\_{t1}$ и $l\_{t1}$ – количество предикатов в первом и втором текстах соответственно. Общая оценка сходства текстов складывается из нормализованных параметров, при этом каждый из них умножается на соответствующий ему коэффициент значимости, полученный по методу линейной регрессии. Коэффициенты приведены на Рисунке 3[[9]](#footnote-9). На нем также показано, что при сравнении текстов наиболее значимыми являются параметры, отражающие полное сходство тех или иных полей. Обучение происходило на 2/3 корпуса изложений, для которых имелись человеческие оценки сходства, 1/3 использовалась для проверки результатов. В целях дополнительной проверки каждый текст тестовой подвыборки был также сравнен сам с собой. При этом обнаружилась незначительная погрешность вычислений с использованием коэффициентов значимости, описанных выше: в некоторых случаях наблюдались отклонения от 1 как в большую, так и в меньшую сторону. Было принято решение ими пренебречь и приравнивать к 1 показатели, значение которых больше 0,8.

Таким образом, оценка сходства текстов считается по формуле:

$$Sim\left(T\_{1},T\_{2}\right)=2 ∙\frac{a\_{1}∙x\_{1}+a\_{2 }∙x\_{2}+a\_{3}∙x\_{3}+a\_{4}∙x\_{4}+a\_{5}∙x\_{5}+a\_{6}∙x\_{6}}{l\_{t1}+l\_{t2}}$$

где $T\_{1},T\_{2}$ – сравниваемые тексты, $a\_{i=0,…,6}$ и $x\_{i=0,…,6}$ – параметры и соответствующие им коэффициенты значимости, $l\_{t1}$ и $l\_{t1}$ – количество предикатов в первом и втором текстах соответственно.

## 3.5. Вычисление сходства новостных текстов. Оценка работы алгоритма

После получения хороших результатов работы алгоритма на корпусе изложений мы перешли к сравнению новостных текстов. Результаты сравнения всех текстов корпуса друг с другом приведены в Таблице 10 (см. Приложение).

Для оценки работы алгоритма была отобрана подвыборка, в которую вошли 15 статей объемом до 300 словоупотреблений. Результаты сравнения этих текстов между собой приведены в Таблице 5: фиолетовым цветом выделены максимальные показатели (больше или равные 0,35), красным – минимальные (меньше или равные 0,1), более сильным тоном обозначены абсолютные максимум и минимум (0,56 и 0,02 соответственно). Оценка производилась вручную и с помощью программы ROUGE (использовалась адаптированная для Python версия [45]).

Таблица 5. Оценки сходства текстов тестовой подвыборки, полученные с помощью нашего алгоритма



Ручная оценка осуществлялась шестью независимыми людьми. Им предоставили подвыборку статей и предложили оценить сходство каждой пары текстов по пятибалльной шкале:

* -2 – тексты совсем не похожи;
* -1 – не похожи;
* 0 – нельзя точно определить, похожи тексты или не похожи;
* 1 – похожи;
* 2 – очень похожи.

Все оценки заносились в таблицу, при этом если разные люди ставили одной и той же паре текстов разные оценки одного знака (включая ноль как для положительных, так и для отрицательных), высчитывался средний показатель, а если оценки различались по знаку, ставился ноль. Результаты приведены в Таблице 6, где зеленым цветом выделены те пары текстов, для которых совпали оценки всех испытуемых, красным – минимальные оценки (меньше -1), фиолетовым – максимальные (больше или равные 1), тоном обозначены абсолютные максимум и минимум (2 и -2 соответственно), для которых также совпали оценки всех испытуемых.

Таблица 6. Человеческие оценки сходства текстов подвыборки



Программа ROUGE предназначена для оценки качества автоматического реферирования и машинных переводов [Lin 2004]. Она сравнивает резюме/аннотации или переводы, сгенерированные машиной, с соответствующими образцами, созданными человеком. Однако, во многих работах ROUGE используется в качестве инструмента для оценки работы того или иного алгоритма по сравнению двух текстов, созданных человеком.

В программе существует три типа оценок: ROUGE-N, ROUGE-S и ROUGE-L. ROUGE-N (обычно берутся ROUGE-1 и ROUGE-2) показывает количество пересечений n-грам в рассматриваемых текстах; ROUGE-L вычисляет самую длинную совпавшую последовательность слов; ROUGE-S рассматривает совпадение скип-грам (n-грам слов, которые могут стоять непосредственно рядом или быть отделены друг от друга определенным количеством слов, в n-граммах не отображенных). При этом программа не лемматизирует словоформы, а берет их в той форме, в какой они встречаются в текстах. Для каждого типа оценки вычисляются три параметра: точность, полнота и F-мера.

Для оценки представленного в данной работе алгоритма сравнения были подсчитаны показатели F-меры для ROUGE-1, ROUGE-2 и ROUGE-L. При этом, поскольку одним из требований программы ROUGE является одинаковое количество строк в сравниваемых файлах, из текстов тестовой выборки было удалено деление на абзацы. Полученные результаты представлены в Таблицах 7-9; фиолетовым цветом выделены максимальные оценки ($\geq 0,35$), красным – минимальные ($\leq 0,1)$, тоном обозначены абсолютные максимум и минимум.

Таблица 7. Показатели ROUGE-1



Таблица 8. Показатели ROUGE-2



Таблица 9. Показатели ROUGE-L



Очевидно, что мера ROUGE-2 показала плохие результаты и в данном случае для оценки непригодна. ROUGE-1 и ROUGE-L в целом соотносятся как с человеческими оценками, так и с показателями сходства, полученными при помощи нашего алгоритма, однако их абсолютные максимумы не совпадают: ROUGE посчитала наиболее схожими тексты, обозначенные как dc\_gazeta\_02 и dc\_gazeta\_05 (оценка 0,46), тогда как люди и наш алгоритм выделили пару sc\_gazeta\_02 и sc\_lenta\_01 (оценки 2 и 0, 56 соответственно). Минимальные показатели соотносимы во всех четырех случаях: строка и столбец, соответствующие тексту el\_freedom\_01 (единственному тексту про выборы президента РФ в данной подвыборке) выделяются полностью, за исключением меры ROUGE-1, согласно которой два текста (dc\_gazeta\_01 и dc\_gazeta\_05) при сравнении с данным получили оценки, равные 0,11.

Таким образом, как видно из Таблиц 5-9, алгоритм, представленный в данной работе, выдает справедливые показатели сходства, наиболее близкие к человеческому восприятию текстов.

# Заключение

В данной работе представлен новый семантико-синтаксический алгоритм сравнения текстов на русском языке, реализованный на Python. В качестве теоретического обоснования алгоритма описаны общие положения построения предложения в русском языке и выделяемые в нем синтаксические структуры различных уровней, а также приведены описания существующих программ по сравнению предложений для различных языков. Кроме того подробно рассмотрены семантические, словообразовательные и семантико-грамматические отношения, связывающие лексемы русского языка.

Разработанный нами алгоритм состоит из нескольких модулей. Первый модуль осуществляет предобработку сравниваемых текстов: они представляются в виде xml-файлов, в которых сохраняется исходное деление на абзацы, для каждого слова указывается его лемма и грамматический тег (полученный с помощью инструмента Pymorphy2 и приведенный к виду, облегчающему последующую обработку), а для каждого знака препинания – терминальность, нетерминальность или неоднозначный статус.

Второй модуль выделяет в полученных файлах чанки, для которых сохраняется вся грамматическая информация о входящих в них словах, а также отдельно отмечается главное слово.

На следующем этапе чанки объединяются в топологические поля: принадлежность к полям подчинительных союзов/союзных слов (Coordination), подлежащих (NomNoun) и сказуемых (FinVerb) определяется по главному слову, остальные поля (LeftNounRem, FreeDet, RightRem, UnableToDefine) выделяются на основании порядка слов в предложении и наличии в нем полей из первой группы.

После выделения топологических полей тексты представляются в виде списка реальных и мнимых предикатов, для каждого из которых содержится информация о формирующем его поле сказуемого (при наличии, иначе указывается «None»), соответствующем поле подлежащего (опять же при наличии), а также список полей распространителей и матрица, отражающая наличие того или иного поля.

Последний пятый модуль осуществляет сравнение текстов по этим предикатам. Подсчитываются значения 6 параметров: $x\_{1}$ – число полностью совпавших предикатов, $x\_{2}$ – сумма показателей сходства частично совпавших предикатов, $x\_{3}$ – число полностью совпавших подлежащих, $x\_{4}$ – сумма показателей частично совпавших подлежащих, $x\_{5}$ – число полностью совпавших распространителей (в полях LeftNounRem, FreeDet, RightRem, UnableToDefine), $x\_{6}$ – сумма показателей частично совпавших распространителей. Каждый из них умножается на соответствующий ему коэффициент значимости, а затем складывается с остальными. Полученная сумма нормализуется по формуле $\frac{2∙x}{l\_{t1}+l\_{t2}},$ где *x* – показатель, $l\_{t1}$ и $l\_{t1}$ – количество предикатов в первом и втором текстах соответственно.

Данный алгоритм был применен для сравнения друг с другом текстов из корпуса новостных статей. Оценка качества его работы осуществлялась вручную и с помощью программы ROUGE (использовались значения F-меры для ROUGE-1 и ROUGE-L) на материалах подвыборки статей, по объему не превышающих 300 словоупотреблений. Анализ полученных данных показал, что наш алгоритм, при сохранении высокой точности и справедливости оценки, представляет показатели сходства, близкие человеческому восприятию текстов, а значит, отражающие реальное положение дел. Кроме того выбранный способ нормализации показателей позволяет сравнивать тексты, сильно различающиеся по объему, без перекоса числовых значений и без ущерба для качества оценки.

Практическая значимость исследования, проведенного в рамках данной работы, заключается в выработке метода выделения в текстах значимых семантико-синтаксических фрагментов, которые упрощают разбор линейных синтаксических структур, и подборе для них весовых коэффициентов, позволяющих получить справедливую оценку сходства при сопоставлении. В дальнейшем представленный нами алгоритм сравнения может быть использован при создании параллельных одноязычных корпусов парафразов и упрощений, для оценки точности автоматического реферирования и машинного перевода а также в программах, нацеленных на обнаружение плагиата. В последнем случае наш алгоритм способен значительно улучшить качество работы антиплагиатных программ, поскольку он не чувствителен к замене слов на синонимы, которая является одним из самых распространенных способов обхода запрета на прямое цитирование без указания авторства.

# Список литературы

1. Азарова, И. В. RussNet как база представления лексического блока в системах автоматизированного анализа текстов: пособие по спецкурсу. – СПб.: СПбГУ. РИО. Филологический факультет, 2015.
2. Азарова И. В. Выявление лексикализованных понятий в RussNet с использованием контекстной информации из корпуса текстов //XXXIII международная филологическая конференция. – №. 25. – С. 3-10.
3. В. А. Белошапкова, Е. А. Брызгунова, Е. А. Земская и др.. Современный русский язык: Учеб. для филол. спец. ун-тов.— М.: Высш. шк. 1989
4. Евсеева И.В., Лузгина Т.А., Славкина И.А., Степанова Ф.В. Современный русский язык: Курс лекций. – Сибирский федеральный ун-т. Красноярск, 2007.
5. Лопатин В.В., Иванова О.Е., Нечаева И.В., Чельцова Л.К. Русский орфографический словарь. – М., 2004. URL: <http://dazor.narod.ru/russkie/slovari/orfograficheskij/orfograficheskij-spisok-sokraschenij.htm>
6. Новикова А.Ю. Семантико-синтаксический анализ глагольных контекстов на основе морфологической дистрибуции. – Магистерская диссертация. СПб.: СПбГУ. Филологический факультет, 2017.
7. Теньер Л. Основы структурного синтаксиса. – М., 1988.
8. Шведова Н.Ю. Русская грамматика. Том II. Синтаксис. – М.: «Наука», 1980.
9. Abney S.P. Parsing by Chunks. In Principle-Based Parsing: Computation and Psycholinguistics, 1991. P. 257–278.
10. Achananuparp P., Hu X., Xiaohua Z., Xiaodan Z. Utilizing Sentence Similarity and Question Type Similarity to Response to Similar Questions in Knowledge-Sharing Community. In WWW 2008 Workshop on Question Answering on the Web, Beijing, China. 2008.
11. Azarova I.V. RussNet as a Computer Lexicon for Russian. In Intelligent Information Systems 2008. P. 447–456.
12. Brill E. A Simple Rule-Based Part of Speech Tagger. In Proceedings, Third Conference on Applied Natural Language Processing, ACL, Trento, Italy, 1992.
13. Brill E. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging // Computational Linguistics. Vol. 21. 1995. P. 543-565.
14. Brill E. Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging //Natural Language Processing Using Very Large Corpora. Somerset (NJ): ACL, 1995. Pp. 1-13.
15. Cheung J.C.K., Penn G. Topological Field Parsing of German // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, 2009. P. 64-72.
16. Dong Z.D., Dong Q. HowNet 1999 [EB/OL]. 2014. URL: <http://www.keenage.com>
17. Dao T.N., Simpson T. Measuring similarity between sentences. 2005. URL: <http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx>
18. Erkan G., Radev D. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research Vol. 22. 2004. P. 457–479.
19. Frey W. Notes on the syntax and the pragmatics of German Left Dislocation. In H. Lohnstein and S. Trissler, editors, The Syntax and Semantics of the Left Periphery. Berlin. 2004. P. 203–233.
20. Hatzivassiloglou, V., Klavans, J., and Eskin, E. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In Proceeding of Empirical Methods in natural language processing and Very Large Corpora. 1999.
21. Hoehle T. Der Begriff ‘Mittelfeld’, Anmerkungen ̈uber die Theorie der topologischen Felder. In Akten des Siebten Internationalen Germanistenkongresses, Gottingen. 1986. P. 329–340.
22. Hoehle T.N. Topologische Felder. Ph.D. thesis, Koln. 1983.
23. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, 2015. P. 320-332.
24. Lang J., Liu T., Li S. LTP: an XML-based open language technology platform. In Proceedings of the 25th Annual Academic Conference of Chinese Information Processing Society of China, 2006.
25. Li L., Hu X., Hu B.-Y., Wang J., Zhou Y.-M. Measuring Sentence Similarity from Different Aspects. In Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding. 2009. P. 2244-2249.
26. Li X., Li Q. Calculation of Sentence Semantic Similarity Based on Syntactic Structure // Mathematical Problems in Engineering, 2015. P. 1-8.
27. Li Y., McLean D., Bandar Z. A., O’Shea J. D., Crockett K. Sentence similarity based on semantic nets and corpus statistics. In IEEE Transactions on Knowledge and Data Engineering Vol. 18. 2006. P. 1138–1150.
28. Lin C.-Y. ROUGE: A package for automatic evaluation summaries. In Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain. 2004. P. 74–81.
29. Liu X., Zhou Y. Zheng R. Measuring Semantic Similarity within Sentences. In Proceedings of ICMLC-2008 Conference, Kunming. 2008.
30. Liu X., Zhou Y., Zheng R. Sentence Similarity based on Dynamic Time Warping. In Proceedings of International Conference on Semantic Computing. 2007. P. 250-256.
31. Mandreoli, F., Martoglia, R., Tiberio, P. A syntactic approach for searching similarities within sentences. In Proceeding of International Conference on Information and Knowledge Management. 2002. P. 656–637.
32. Meurers W.D. On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. Lingua. Forthcoming. 2002.
33. Mihalcea, R., Corley, C., Strapparava, C. Corpus-based and knowledge-based measures of text semantic similarity. In Proceeding of the Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference. 2006.
34. Papineni K., Roukos S., Ward T., Zhu W. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002. P. 311–318.
35. Parris S.R. Adolescent Literacy, Field Tested: Effective Solutions for Every Classroom. International Reading Assoc. 2009.
36. Veenstra J., Müller F., Ule T. Topological Field Chunking for German // Proceedings of the 6th Conference on Natural Language Learning. Stroudsburg, PA, USA, 2002. P. 1-7.

# Электронные источники

1. Викисловарь

URL: <https://ru.wiktionary.org/wiki/>

1. ГОСТ 7.12-93. Сокращения слов в русском языке

URL: <https://studfiles.net/preview/5229454/>

1. Документация Pymorphy2

URL: <https://pymorphy2.readthedocs.io/en/latest/>

1. Национальный корпус русского языка

URL: <http://www.ruscorpora.ru/>

1. Радио «Свобода»

URL: <https://www.svoboda.org/>

1. РИА Новости

URL: <https://ria.ru>

1. Gazeta.ru

URL: <https://gazeta.ru>

1. Lenta.ru

URL: <https://lenta.ru>

1. OpenCorpora

URL: <http://opencorpora.org/?page=downloads>

1. PM-TXM

URL: <https://github.com/vintagentleman/PM-TXM>

1. Rouge

URL: <https://github.com/pltrdy/rouge>

1. RussNet

URL: <http://project.phil.spbu.ru/RussNet/index_ru.shtml>

1. Simple Wikipedia

URL: <https://simple.wikipedia.org/wiki/Main_Page>

# Приложение

## Код представленного в данной работе алгоритма

URL: <https://github.com/laleopal/TEXTCOMP>

## Обозначения компонентов структурных схем

* $V\_{fin}$— спрягаемая форма глагола;
* $V\_{3s}$ — спрягаемый глагол в форме 3 л. ед. ч.;
* $V\_{3p}$ — спрягаемый глагол в форме 3 л. мн. ч.;
* *Inf* — инфинитив;
* $N\_{x=1,…,6}$ — существительное, цифры от 1 до 6 обозначают падежи: 1 — им. п., 2 — род. п., 3 — дат. п., 4 — вин. п., 5 — тв. п., 6 — предл. п.;
* *Adj* — прилагательное;
* *Adj\* —* краткое прилагательное;
* *Pron* — местоимение;
* *Adv* — наречие;
* *Adv­o* — предикативное наречие на о;
* *Praed* — предикатив;
* *Part* — причастие;
* *Part\* —* краткое причастие;
* $Praed \_{part}$— причастный предикатив;
* *interj* — междометие;
* *neg* — отрицание;
* *cop* — связка;
* *quant* — квантификативное (количественное) значение.

## Корпус новостных статей и параллельный корпус изложений

URL: <https://github.com/laleopal/TEXTCOMP/tree/master/inpt>

## Список парафразов, составленный на материале корпуса изложений

• сложноподчиненное предложение, в котором подлежащие в главном и придаточном совпадают, и простое предложение с однородными сказуемыми («Мы так привыкли к этому образу, что иного Ломоносова не представляем» – «Привыкли уже и не представляем иначе», Т1);

• сказуемое с развернутым дополнением, выраженным существительным с прилагательным, и сказуемое с обстоятельством, выраженным наречием («иного Ломоносова не представляем» – «не представляем иначе», Т1);

• развернутое обстоятельство, выраженное двумя существительными с предлогами в родительном и творительном падежах соответственно («во время наблюдения за Венерой», предл. 20), придаточное предложение с союзом «когда» («когда Ломоносов изучал планету Венеру», Т5) и деепричастный оборот («смотря из телескопа на Венеру», Т23);

• развернутое дополнение, выраженное двумя существительными в винительном и родительном падежах и местоимением с предлогом в предложном падеже («существование на ней атмосферы», предл. 20), и придаточное предложение, вводимое союзом «что» («что на ней может быть атмосфера», Т23);

• составное именное сказуемое с дополнением-инфинитивом («легче разглядеть», предл. 7) и составное глагольное сказуемое с обстоятельством-наречием («можно легко увидеть», Т10);

• развернутая группа сказуемого с обстоятельством-наречием и дополнением-существительным («и не так уж много надо воображения», предл.12) и группа сказуемого с местоимением в дательном падеже и дополнением-существительным с определением-прилагательным («нам не нужно большого воображения», Т13);

• дополнение-существительное с развернутыми определениями, выраженными причастным оборотом и двумя существительными («идущего с веслом парня по имени Михаил», предл.12), и дополнение-существительное с определениями, выраженными прилагательным и существительными: в родительном падеже без предлога и в творительном падеже с предлогом («простого парня Михаила с веслом», Т13);

• существительное с определением-прилагательным («по русской истории», предл.18) и существительное с определением, выраженным существительным в родительном падеже («по истории России», Т13);

• развернутое определение, выраженное глагольным инфинитивом с существительным с предлогом в винительном падеже и существительным с определением-прилагательным в творительном падеже («пройти на восток северными морями», предл. 19), и развернутое определение, выраженное глагольным инфинитивом с существительным с предлогом в родительном падеже и существительным с предлогом и определением-прилагательным в дательном падеже («дойти до востока по северным морям», Т13);

• простые предложения с составными именными сказуемыми («Ломоносов был верным и преданным сыном России. Он был великим гражданином Отечества», предл.23-24) и назывные предложения («Верный и преданный сын России. Великий гражданин Отечества», Т13);

• сказуемое-глагол с обстоятельством, выраженным прилагательным в именительном падеже («первый указал», предл.19) и сказуемое-глагол с обстоятельством-прилагательным в творительном падеже («первым указал», Т17);

• сказуемое-глагол с дополнением-существительным с предлогом в винительном падеже («указал на возможность», предл.19) и сказуемое-глагол с дополнением, выраженным существительным в винительном падеже без предлога («указал возможность», Т17);

• бессоюзное придаточное изъяснительное предложение («он был первым нашим университетом», предл.27) и придаточное, вводимое союзом «что» («что он был первым нашим университетом», Т21);

• составное именное сказуемое, выраженное глаголом и существительным в творительном падеже с определениями-прилагательными («был первым нашим университетом», предл.27), и именное сказуемое, выраженное существительным и определениями-прилагательными в именительном падеже («наш первый университет», Т22);

• определение, выраженное существительными в родительном падеже («деревни Ломоносово», предл.5), и обстоятельство места, выраженное существительными с предлогом в предложном падеже («в деревне Ломоносово», Т6, Т7; «в деревне Ломоносова», Т19);

• простое предложение с причастным оборотом («В этом лице легче разглядеть русского парня, вырастающего на краю деревянной России», предл.7), сложноподчиненное предложение с придаточным изъяснительным («Говорят, что парень вырос в деревянной России», Т1) и сложноподчиненное предложение с придаточным определительным («В лице он больше похож на русского парня, который проживал на краю деревянной России», Т4);

• полное простое предложение с составным именным сказуемым («Он был первым нашим поэтом и первым физиком», предл.14) и назывное предложение («Первый поэт и физик», Т1);

• определение, выраженное существительным с прилагательными с предлогом в предложном падеже («в белом завитом парике», предл.2), и определение, выраженное существительным и прилагательным с предлогом в творительном падеже («с белым париком», Т3, Т7);

• простое распространенное предложение «При этом слове мы сразу же вспоминаем человека в белом завитом парике, румяного и полнолицего» (предл.2), сложное предложение «Вспоминаем Ломоносова: толстощекий румяный мужчина» (Т15) и простое предложение с составным именным сказуемым («Ломоносов – это человек в белом парике, румянолицый», Т22);

• сочетание «при этом слове» (предл.2), местоимение «его» (Т4), сочетание «при любом слове о Ломоносове» (Т23), деепричастный оборот «слыша это имя» (Т5) и придаточное предложение («когда мы слышим эту фамилию», Т13; «когда говорят о Ломоносове», Т21);

• дополнение-существительное в родительном падеже («(вспоминаем) человека», предл.2), дополнение-существительное с союзом «как» («(вспоминаем) как человека», Т4) и дополнение, выраженное существительным с предлогом в предложном падеже («(вспоминаем) о человеке», Т7);

• группа сказуемого «так привыкли к этому образу» (предл.3), в которую помимо самого глагола вошли усилительная частица и существительное и прилагательное с предлогом в дательном падеже, и группа «его знаем таким» (Т4), состоящая из глагола, местоимения в родительном падеже и прилагательного в творительном падеже;

• простое глагольное сказуемое («стоишь», предл.5) и составное глагольное сказуемое («можно стоять», Т4);

• полное простое предложение с составным именным сказуемым, выраженным глаголом и существительными с прилагательными («Ломоносов был верным и преданным сыном России», предл.23), полное простое предложение с составным именным сказуемым, выраженным глаголом и краткими прилагательными («России Ломоносов был верен и предан», Т4; «Ломоносов был очень верен и предан своей Родине», Т5) и полное простое предложение с составным именным сказуемым, выраженным глаголом с полными прилагательными и существительным («Ломоносов был верным и преданным России», Т6; «Ломоносов был верным и преданным своему Отечеству», Т23);

• сложноподчиненное предложение «И не так уж много надо воображения, чтобы представить себе идущего с веслом парня по имени Михаил» (предл.12) и полное простое «В этой русской деревне мы бы могли представить себе Михаила, идущего с веслом» (Т23);

• сказуемое с дополнением, выраженным местоимением с предлогом в предложном падеже («говорил о нем», предл.27), и сказуемое с дополнением, выраженным местоимением в родительном падеже («называл его», Т10);

• простое предложение «А ведь был и без парика Ломоносов» (предл.4), предложение с обособленным определением «А ведь был и другой Ломоносов, без парика» (Т13) и предложение «Но он не всегда был таким» (Т21);

• обстоятельство места, выраженное существительным и прилагательным с предлогом в творительном падеже («перед этим портретом», предл.5; «над этим портретом», Т18) и обстоятельство места, выраженное существительным и прилагательным с предлогом в родительном падеже («у этого портрета», Т13);

• обстоятельство образа действия, выраженное существительными с предлогом в винительном падеже и без предлога в родительном падеже («в окошко музея», предл.11), обстоятельство места, выраженное существительным с предлогом в творительном падеже («за окном», Т13) и обстоятельство образа действия, выраженное существительными с предлогом в винительном падеже и с предлогом в предложном падеже («из окошка в музее», Т17);

• определение-существительное с предлогом в предложном падеже («(лед) на реке», предл.11) и определение-существительное в родительном падеже («(лед) реки», Т13);

• сочетание «во время наблюдения за Венерой» (предл.20) и деепричастный оборот «увидев Венеру» (Т13);

• сочетание «существование на ней атмосферы» (предл.20) и придаточное определительное «что на ней присутствует атмосфера» (Т13);

• предложение «место ему в ряду титанов эпохи Возрождения» (предл.21) и безличное предложение «его можно причислить к титанам эпохи Возрождения» (Т13);

• группа сказуемого, состоящая из наречия, глагола и существительного в родительном падеже («высоко ценил Ломоносова», предл.26), и сказуемое с дополнением-местоимением («им восхищался», Т13);

• группа сказуемого, состоящая из глагола, глагольного инфинитива и местоимения с предлогом в винительном падеже («умел за себя постоять», предл.25), и группа сказуемого, помимо вышеуказанного распространенная еще усилительной частицей «и» и еще одним местоимением («и сам мог за себя постоять», Т19).

## Таблица 10. Оценка сходства текстов из новостного корпуса

Цветовые обозначения: красный – минимальные оценки ($\leq 0,1$), фиолетовый – максимальные ($\geq $0,35), тоном выделены абсолютные минимум и максимум, зеленый – сравнение текста самого с собой.



1. Полный список и расшифровка используемых символов приведены в Приложении к данной работе. [↑](#footnote-ref-1)
2. Ссылка на код находится в Приложении [↑](#footnote-ref-2)
3. Примеры тегсетов см. на Рисунке 2. [↑](#footnote-ref-3)
4. Список сокращений был составлен по материалам «Русского орфографического словаря» [5] и ГОСТа сокращений в русском языке [35]. [↑](#footnote-ref-4)
5. Перечень союзов и союзных слов был составлен с опорой на Викисловарь [37]. [↑](#footnote-ref-5)
6. При сравнении как главных, так и зависимых слов не учитываются местоимения, союзы и вспомогательный глагол «быть». [↑](#footnote-ref-6)
7. Для полей распространителей накапливаются 2 показателя, отражающие сходство главных и зависимых слов соответственно. [↑](#footnote-ref-7)
8. Числовые значения подобраны с опорой на практический опыт и работу [26]. [↑](#footnote-ref-8)
9. При вычислении сходства текстов данные коэффициенты округлялись до сотых. [↑](#footnote-ref-9)