ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ  
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ  
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

На правах рукописи

Буйлова Надежда Николаевна

**Поверхностный синтаксис глагола**

**как маркер литературных формул**

Резюме   
диссертации на соискание ученой степени  
кандидата филологических наук (PhD)

Научный руководитель   
кандидат филологических наук, профессор  
Ляшевская Ольга Николаевна

Москва, 2021

**Общая характеристика исследования**

В диссертационной работе рассматриваются глагольные конструкции, присущие определенным жанрам литературы (любовному роману, детективу, научной фантастике и фэнтези). С помощью метода автоматического распознавания жанра выявляются конструкции, характерные для этих жанров. Полученные конструкции анализируются количественно и качественно. Доказывается, что высокоуровневые признаки текста могут служить признаками для машинного классификатора.

Диссертация выполнена в рамках теоретической парадигмы Грамматики конструкций. В рамках направления исследования глагольных конструкций значимы работы Ч. Филлмора, А. Голдберг, М. Томазелло, Х. Боаса, Л. Микаэлис и др. На материале русского языка исследования семантико-синтаксических свойств глагола в функциональной и когнитивной традиции проводились Ю. Д. Апресяном, И. А. Мельчуком, В. С. Храковским, Е. В. Падучевой, Г. И. Кустовой, В. Ю. Апресян, Е. В. Рахилиной, Ю. Л. Кузнецовой, С. С. Саем и др. В части проблематики определения литературной формулы данная диссертация опирается как на работы литературоведов и культурологов (Дж. Кавелти, А.-М.Бойе, С. Бордони, Т.Г. Скребцовой, Н.М. Марусенко и др.), так и на работы, выполненные в компьютерно-лингвистической традиции (Б. Кеслера, Ю. Карлгрена, Д. Каттинга, Х. Шутце, Ф. Себастиани и др.).

Работа основана на материале глагольных конструкций русского языка. **Цель данной работы** – выявление специфики употребления конструкций в близких жанрах литературы.

Для достижения цели работы необходимо решить следующие **задачи**:

— охарактеризовать исследуемые жанры, определить их специфику

— сформулировать принципы отбора текстов для проведения экспериментов,

— адаптировать основные понятия грамматики конструкций к задаче определения жанра с использованием синтаксической информации,

— определить границы вариативности глагольных конструкций в близкородственных жанрах,

— выделить группы глаголов, вносящих наибольший вклад в различение жанров,

— проанализировать эти глаголы и их конструкционные свойства.

В центре внимания исследования находятся характерные синтаксические особенности современной массовой литературы. Эти специфические характеристики не рассматривались в комплексе в исследованиях по синтаксису и машинному обучению, что обусловливает – вкупе с их потенциальной значимостью для уточнения тегов жанра – **актуальность** настоящей работы.

Интерпретация языковых данных осуществляется с применением различных лингвистических **методических приемов и процедур**, в частности, автором работы создан алгоритм, позволяющий выделить глагольные конструкции из художественной литературы, составлены списки жанроспецифичных глаголов, проведены эксперименты по машинному обучению классификатора.

**Теоретическая значимость** исследования заключается в рассмотрении синтаксической конструкции как важного стилистического маркера. Разделение глаголов на категории по количеству заполненных валентностей, рангу конструкций заданного порога, а также выделение кластеров, характерных для каждого жанра, может задать вектор развития стилометрических исследований литературных формул.

**На защиту выносятся следующие положения**:

1. Жанроспецифичные глагольные конструкции (конструкции-маркеры) – это конструкции глагола с его синтаксическими зависимыми, употребление которых в одном из жанров значимо преобладает относительно их употребления в других жанрах.
2. Задача автоматического определения микрожанра текста может решаться не только за счет низкоуровневых лингвистических признаков (длина слова, предложения и т.д.), но и на уровне лексики и лексических конструкций.
3. Частотное распределение конструкций отдельного глагола в корпусе неравномерно: в среднем две-три самые частотные конструкции покрывают 50% употреблений глагола.
4. Метрики значимости и важности признаков, используемых в алгоритмах автоматической классификации, могут быть использованы для выявления жанроспецифичных глагольных конструкций.
5. Выделяется несколько групп глагольных конструкций (полные VS неполные, аргументные VS c семантически необязательными модификаторами, конструкции отдельных тематических групп глаголов), вариативность которых маркирует жанр.
6. Анализ различий в употреблении глагольных конструкций в разных жанрах позволяет связать их с жанровыми признаками более высокого уровня (тема, тренд, авторские ограничения на лексику).

**Практическая значимость** исследования основана на особенностях современного подхода к определению жанра текста. В современном машинном обучении чаще всего исследуются модели, основанные на статистических признаках и слабо связанные с высокоуровневыми параметрами текста, такими, как строение предложения или словосочетания; рассмотрение глагольной конструкции в таком аспекте представляется нам важным шагом.

Научная фантастика, детективы, любовные романы и фэнтези – представители формульной литературы, которые могут служить материалом для разработки метода автоматического определения микрожанра. В рамках исследования создан корпус произведений массовой литературы, который может быть использован как набор данных для оценки качества алгоритмов определения микрожанра текста.

**Актуальность и значимость** работы заключается в разработке алгоритма, позволяющего улучшить метаразметку корпусов русского языка. Сейчас большая часть корпусов имеет множество слоев разметки (морфологическую, синтаксическую, семантическую), однако задача определения метатегов, связанных с такими данными, как локация или время создания текста, большей частью ложится на плечи не программистов или лингвистов, а литературоведов и историков. Несмотря на достаточную логичность подобного положения дел, некоторые метатеги вполне могут быть проставлены автоматически – к примеру, жанр текста, некоторые характеристики стихотворных текстов и пр. Разработанный в рамках данной диссертации алгоритм определения жанра текста может быть применен в разметке корпусов; особо примечательно, что алгоритм способен различать близкородственные жанры. Предполагается, что этот подход может быть использован для определения не только жанра, но и других характеристик текста, к примеру, авторства и времени написания.

Научная новизна работы заключается в том, что конструкционный потенциал лексики до нашего исследования не рассматривался как потенциально значимый для определения жанра текста. Большая часть работ подобного рода строятся на более низкоуровневых маркерах, таких как частеречные характеристики текста, распределение служебных слов или буквенные биграммы.

**Апробация работы.** Основные результаты работы были представлены на конференциях «Колмогоровские чтения» (Москва 2017), ARANEA (Братислава 2018), опубликованы в журналах НТИ сер.2, CLLS’18 и Вестнике ПСТГУ сер.III Филология, 2021 вып. 66 стр. 11-23.

**Структура работы**. Работа содержит введение, пять основных глав, заключение и список литературы. В первой главе излагаются теоретические основания использования глагольных конструкций как признака для машинного классификатора. В частности, в разделе 1.1 рассматриваются теоретические основы грамматики конструкций, в разделе 1.2. описываются подходы к определению жанра текста, описываются основные работы и подходы к изучению синтаксиса, а в разделе 1.3 рассматриваются исследования по машинному обучению в различении жанров. Во второй главе описаны используемые данные и ресурсы. Корпуса, используемые в исследовании, рассмотрены в разделе 2.1. В третьей главе рассматривается метод обработки данных, его возможности и ограничения. Признаки для машинного обучения и алгоритмы рассматриваются в разделах 3.2 и 3.3. В четвертой главе описываются проведенные эксперименты – машинное обучение с применением низкоуровневых (раздел 4.1) и высокоуровневых (раздел 4.2) признаков. В пятой главе представлен анализ полученных лингвистических данных. В заключении подводятся итоги исследования.

**Основное содержание работы.**

Изучение поверхностного синтаксиса на статистических данных стало возможным после появления больших корпусов и универсальных зависимостей. При этом традиционно изучается скорее нормативный язык, язык «большой литературы», а детская, массовая литература, нон-фикшен остаются в ее тени. В нашем исследовании мы обратились к так называемой «формульной литературе» и поставили задачу выяснить при помощи количественных методов, каким образом поверхностный синтаксис подобных произведений может обуславливать их «формульность». Таким образом, **обзор литературы** включает в себя три раздела, которые затрагивают ключевые объекты диссертации: в Разделе 1.1 описаны статьи и монографии, посвященные конструкционной грамматике, в Разделе 1.2 приведен краткий обзор современных взглядов на литературные формулы и сопоставимые феномены, в Разделе 1.3 обсуждаются основные работы по применению машинного обучения для классификации текстов различных жанров.

В нашей работе мы изучаем синтаксис, используя вербоцентричную теорию зависимостей, разработанную Л. Теньером и опубликованную в книге «Основы структурного синтаксиса» [Теньер, 1988]. Теньер первым ввел понятие глагольной валентности и классифицировал глаголы по количеству актантов, которые они могут присоединять, а также описал средства изменения валентной структуры. Теория актантов Теньера стала основой для грамматики конструкций (CxG) Ч. Филмора, А. Голдберг и др. В качестве базовых принципов, разделяемых сегодня всеми направлениями CxG, можно назвать отрицание четкой границы между грамматикой и словарем и отказ от статичного взгляда на язык. Взаимные ограничения «уравнивают в правах» все составляющие системы – именные группы, относящиеся к участникам ситуации, такие же элементы глагольных конструкций, как и сам глагол. Семантика, синтаксис и даже морфология в таких конструкциях подвижны и взаимообусловлены.

В контексте Грамматики конструкций большое значение имеют работы Ю.Д. Апресяна, например, «Экспериментальное исследование русского глагола» [Апресян, 1967], в котором постулируется прямая связь синтаксиса и семантики. В этой работе проанализировано примерно полторы тысячи самых частотных русских глаголов и их конструкций. Описанные конструкции отвечают определенным требованиям, например, они не имеют сирконстантных позиций. Глагольные конструкции охарактеризованы по формальным принципам: строятся трансформационные классы, деревья, иерархические классификации и т. д. Развитием исследований сирконстантов стало открытие, что сирконстант связан с глаголом семантически — он не может по смыслу противоречить никакой части конструкции. Впервые семантическая сочетаемость актанта и сирконстанта была рассмотрена в работе «Сирконстанты в толковании?» [Плунгян, Рахилина, 1990]*,* в которой было показано, что определенные сирконстанты сочетаются с глаголами, описывающими конкретные ситуации (*бежать, резать*), и не сочетаются с ситуациями абстрактными (*портить, мстить*).

При отборе данных мы используем определение Дж. Г. Кавелти из работы «Изучение литературных формул» [Cawelty, 1976]: «формулы – это способы, с помощью которых конкретные культурные темы и стереотипы воплощаются в более универсальных повествовательных архетипах». Для нашего исследования мы будем использовать несколько формул, предложенных Дж. Г. Кавелти: «Приключение», «Романтическая история», «Тайна», «Чужие сущности и состояния». Легко видеть, как рассматриваемые нами разновидности массовой литературы соотносятся с четырьмя формулами Кавелти. Так, формуле «романтическая история» соответствует любовный роман, «тайне» – детектив, «приключению» – фэнтези, последней формуле, «чужим» – научная фантастика.

Изучение литературных формул позволяет не только дистанцироваться от конкретных сюжетных особенностей текста, но и перейти непосредственно к изучению лингвистических особенностей и языковых механизмов. Воспроизводимость сюжетов и миров ожидаемо порождает тематическую предсказуемость лексики и воспроизводимый выбор лексических средств. Стандартизация подобного рода произведений удобна для лингвиста, поскольку позволяет сосредоточиться на языковых особенностях текста, максимально дистанцировавшись от сюжета.

Существует несколько способов изучать массовую литературу. Розовый любовный роман зачастую рассматривается в контексте современной социокультурной парадигмы, изучая причины его популярности и «эволюцию» формальных признаков жанра. Практически все любовные романы на русском языке – калька с англоязычных произведений с минимальным заимствованием современного антуража. Отмечается, что повествование по канону жанра не может вестись от первого лица, а, следовательно, и глаголов в такой видовой форме не будет.

Другой подход, использованный при изучении детективов, связан с исследованием корней двух наиболее крупных подвидов современного детектива: английскому (в центре которого находится расследование, скорее «женский детектив») и американскому (в центре которого находится преследование и физические стычки, скорее «мужской детектив»).

Фэнтези и фантастика – два жанра, имеющие близкие формулы «приключение» + «чуждые существа и состояния» с разными акцентами. По мнению большинства исследователей, особенностью фэнтези, отличающей его от научной фантастики, является принципиальная невозможность воплощения книги в реальность. Это порождает отдельные замечания о возможности построения словаря подобного жанра (который зачастую включать в себя как авторские неологизмы, так и архаизмы). Для микрожанра «космическая боевая фантастика» характерна принципиальная объяснимость мира. Одновременно с этим в выбранном нами поджанре присутствуют черты «приключения» (перестрелки, решение проблем грубой силой) и «чуждых существ и состояний» (действие происходит в космическом пространстве и/или на других планетах (в Солнечной системе или за её пределами) в условном (обычно экзотическом) антураже).

Легко видеть, что в современных исследованиях формульной литературы значительно большее внимание уделяется литературоведческим и культурологическим особенностям произведений, а лингвистические характеристики текста в целом и конструкционный потенциал жанров изучен ограниченно.

Особенности микрожанров могут быть использованы для их автоматического различения. Современные исследования по построению дифференциальных моделей, сбору однородных корпусов и классификации текстов сталкиваются с проблемой обусловленности признаков не только жанровыми особенностями, но и дополнительными факторами (разногласия между аннотаторами при разметке текстов, большое количество жанров и т.д.), которые обуславливают сложности определения границ даже крупных жанров. Современные исследователи стремятся найти новые жанроразличительные признаки, в числе которых – поверхностный синтаксис глагола.

Ранние работы по распознаванию жанра текста применяли методы дискриминантного анализа и логистической регрессии, в том числе с использованием нейросети. В качестве базового описания текстов были использованы частеречные характеристики текста, а также различные меры удобочитаемости. В дальнейшем для определения жанра использовались методы мешка слов и деревьев решений, основанные на частеречной разметке. Другим крупным ответвлением классификации документов стала классификация с использованием HTML-разметки, позволяющей комбинировать квантитативные методы описания самого текста с нетекстовыми элементами разметки гипертекста. Кроме того, следует упомянуть использование синтаксически размеченных корпусов (так называемых «treebank»-ов), позволяющих проводить анализ дискурсивных связей.

В **Главе 2** описаны используемые в работе данные. Они включают подкорпуса текстов художественной литературы: детективов, любовных романов, фэнтези и научной фантастики, каждый из которых содержит более 280 текстов. Тексты отбирались на основе пользовательской разметки жанров; общий объем корпуса составил 104 919 587 токенов (любовных романов – 18 205 059; детективов – 24 038 408; научной фантастики – 30 086 136; фентези – 32 589 984).

Так как теги жанровой атрибуции, проставляемые пользователями открытых источников, могут содержать ошибки, все тексты были просмотрены вручную на соответствие указанному жанру. Кроме того, так как внутрижанровое разнообразие довольно велико (любовные детективы, технофентези и т.д.), наша работа концентрируется на следующих поджанрах:

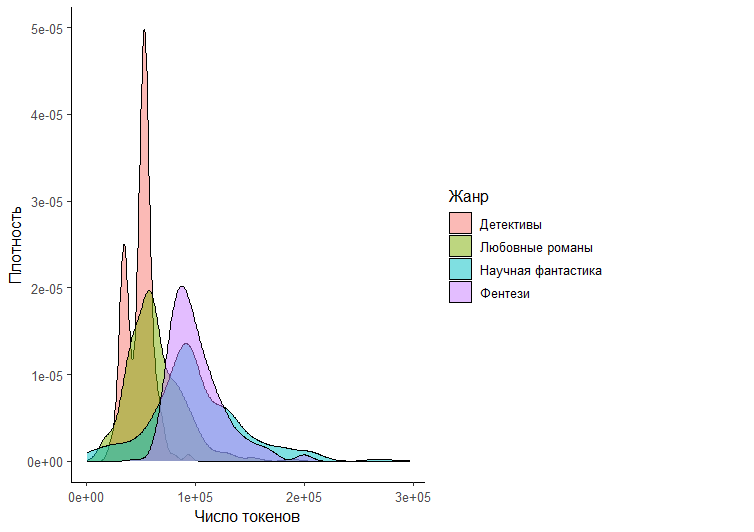
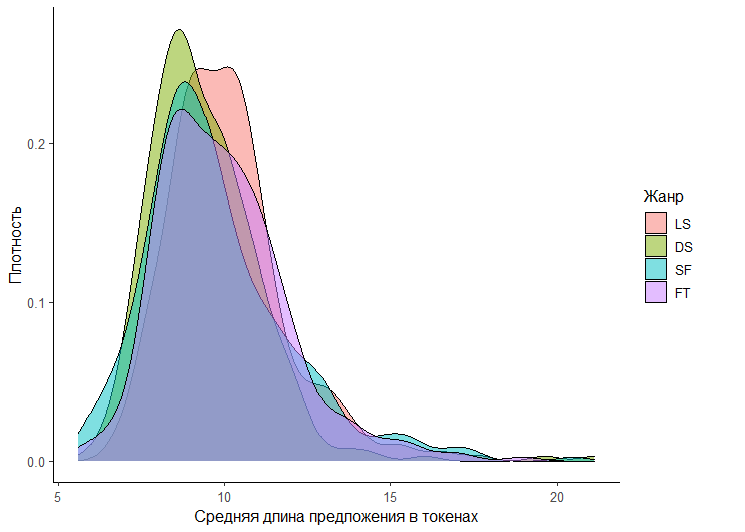
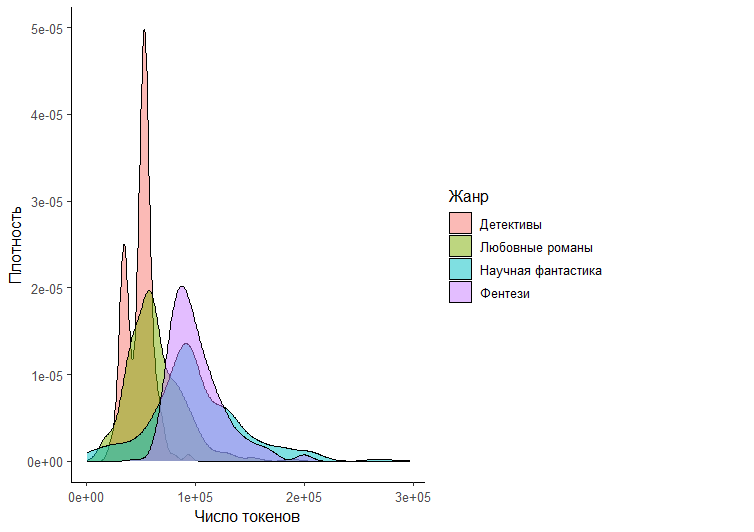
1. «Розовый» любовный роман, сконцентрированный на романтических и эротических описаниях.
2. Крутой детектив – детектив, в который на первый план выходит не умение героя или героини делать логические выводы, как в классическом детективе, а выносливость, умение стрелять или драться врукопашную.
3. «Попаданческая» или боевая фантастика, описывающая столкновения с иными формами жизни. Достижения науки, обуславливающие развитие этого мира, часто не имеют отношения к реальности и подчиняются исключительно воле автора.
4. «Попаданческое» или высокое фентези, основанное на магическом мироустройстве. Герой подобных произведений – неизменно «избранный», обладающий особыми способностями, которые резко выделяют его из толпы.

Далее, привлекалось не более пяти произведений от автора в категории, что, по нашему мнению, позволило существенно снизить влияние индивидуально-авторских особенностей текста в рассматриваемом материале. В результате фильтрации итоговый объем корпуса сократился до 1201 произведения (280 любовных романов, 319 детективов, 304 фентези, 321 научный роман).

Собранные корпусы были размечены при помощи программы автоматического морфологического и синтаксического анализа UDPipe 2.6 в статистической среде R с моделью UD-SynTagRus 2.6. В результате для каждого токена была получена лемма, тег части речи, грамматические признаки леммы, а также дерево синтаксических отношений с вершинами и зависимыми (Рис. 1). Далее проводилась постобработка данных для получения комплексов вида «root-nsubj», «root-ccomp», «root-obl-obl», то есть коллокаций с определенным типом связи.

**Рисунок 1.** Пример разметки UDPipe.

Мы предполагаем, что случайный отбор текстов моделирует реальную ситуацию, сложившуюся в литературной практике, поэтому тексты не нормировались по длине. График плотности распределения подкорпусов по длине текстов представлен на рисунке 2а. Поскольку основной целью эксперимента было обнаружение глагольных конструкций, вносящих наибольший вклад в машинное обучение, признак длины текста в дальнейшем не использовался в экспериментах. На рисунке 2б показано распределение длин предложений в словоформах. Графики существенно пересекаются, поэтому этот признак также не может быть использован для эффективного различения жанров. Выбранные микрожанры также не различались по распределению частеречных тегов. Что касается тегов зависимостей, наиболее частотной связью являлась пунктуация, после которой шли подлежащее и косвенное дополнение. В большинстве случаев показатели по четырем жанрам различались незначительно.



а

б

**Рисунок 2.** Количественные характеристики подкорпусов. **а.** Распределение количества токенов в текстах подкорпусов. **б.** Распределение длин предложений в токенах в подкорпусах.

С точки зрения лексического многообразия корпуса демонстрируют прирост уникальных лексем в соответствии с приростом размера корпуса: научная фантастика имеет самый богатый словарь.

Таким образом, микрожанры достаточно схожи друг с другом по целому ряду параметров: длине предложения, распределению частей речи, лексическому разнообразию. Несколько различаются длины текстов – это может быть объяснено традиционным тяготением научной фантастики к более крупным формам (трилогиям, тетралогиям и т.д.), которые попали в нашу выборку.

В **Главе 3** рассматривается общая схема проведения исследования. Как было показано выше, частные литературные жанры слабо дифференцированы по низкоуровневым признакам, что осложнит работу классификатора. Другой уровень признаков – «мешок слов» – так же мало подходит для классификатора из-за низкого лексического разнообразия жанров массовой художественной литературы. В нашем исследовании низкоуровневые признаки были дополнены глагольными лексемами и глагольными конструкциями, предположительно неодинаково представленными в рассматриваемых поджанрах.

Глагольные конструкции, как было описано выше, извлекались из подкорпусов с помощью сервиса UDPipe. Следует отметить, что разметка синтаксических связей имеет погрешности, однако анализ случайно выбранных из подкорпусов предложений показал, что качество разметки не ниже результатов, показанных на CoNLL 2018 Shared Task.

Для классификации жанров использовались следующие признаки:

1. низкоуровневые признаки (частотный вектор длины словоформ и частотный вектор длины предложений);
2. признаки глагольных лексем (вектор частот глагольных лемм);
3. признаки синтаксических конструкций конкретных глаголов (частотный вектор синтаксических конструкций конкретного глагола).

Частотные вектора строились с использованием метода нормализации частот с помощью метрики LL-score (коэффициента логарифмического правдоподобия). Эта метрика показывает ключевые единицы для каждого поджанра (*убивать* и *бояться* в типичных детективах, *жениться*, *любить* в любовных романах). Глаголы и конструкции, имеющие LL-score выше 3,5, мы называем жанроспецифичными.

Синтаксическая конструкция определяется как набор синтаксических зависимых, реализованных при глаголе в тексте, например, <*превращаться;nsubj;obl*> для предложения *Он превратился в чудовище*. Нами были отобраны шесть типов связей – четыре актантных (обязательных для построения грамматически и семантически правильного предложения) и две сирконстантных (факультативных), а также их сочетания: nsubj (объект, подлежащее); obj (субъект, прямое дополнение); ccomp (комплементарная клауза); xcomp (открытая клауза); obl (косвенное дополнение); advcl (клауза со значением причины, обстоятельства, следствия).

При анализе учитывалось только наличие связей, но не порядок их появления: <*nsubj;obj*> и <*obj;nsubj*> считаются одинаковыми конструкциями. Нами были проанализированы частоты встречаемости конструкций. Определенные синтаксические отношения имеют более высокую частотность в корпусе детективов (<*ccomp;nsubj*>, <*obl;xcomp*>) или любовных романов (<*nsubj;obl;obl*>), что может быть маркером жанра. Однако подобных связей немного, поэтому необходимо сравнивать отдельные глаголы, имеющие в разных жанрах разные конструкции. Для определения таких глаголов мы ввели понятие «конструкция заданного порога» (construction at specific threshold, CST) – это номер конструкции из списка конструкций глагола, отсортированных по убыванию частотности, на которой суммарное количество употреблений глагола превысило N% от общего количества. Полученные нами результаты показали, что, во-первых, количество конструкций глагола, необходимых для пересечения порога, редко изменяется при смене жанра, во-вторых, три наиболее частотных конструкции глагола чаще всего совпадают, но при этом их частотный ранг может меняться.

В **Главе 4** представлены результаты применения алгоритмов машинного обучения для классификации текстов по поджанрам. В нашем исследовании было применено три алгоритма и три набора признаков (Таблица 1).

В нашем исследовании были использованы несколько алгоритмов машинного обучения: наивный баейсовский классификатор, деревья решений и случайный лес. Использовался язык программирования Python 3.1 и модуль os. Настройки гиперпараметров классификатора не проводилось, поскольку нас интересовал вклад лингвистических параметров, а не возможности подстройки алгоритма. Наивный байесовский метод исползовался как отправная точка (baseline), позволяющий оценить прирост качества классификатора. Выбор деревьев решений и случайного леса был обусловлен простотой метода и легкостью интерпретации получаемых результатов.

Были использованы три набора признаков: первый (базовый) включал в себя низкоуровневые признаки (длину слова в символах и предложения в словах, частеречные характеристики), второй использовал также взвешенный список глагольных лексем, составленный при помощи коэффициента подобия, третий добавил к первым двум лексико-синтаксические признаки (список глаголов с конструкциями).

**Таблица 1.** Результаты экспериментов с машинным обучением (P – precision, точность; R – recall, полнота; F – f-score, f-мера).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Наивный байесовский классификатор | | | Деревья решений | | | Случайный лес | | |
|  | P | R | F | P | R | F | P | R | F |
| Базовый набор признаков (длина слов + длина предложения) | 0.43 | 0.47 | 0.43 | 0.47 | 0.47 | 0.46 | 0.5 | 0.53 | 0.51 |
| Базовый набор признаков + леммы (нормализация LL-score) | 0.6 | 0.58 | 0.57 | 0.73 | 0.8 | 0.79 | 0.82 | 0.85 | 0.83 |
| + глаголы с конструкциями (нормализация LL-score) | 0.66 | 0.66 | 0.66 | 0.86 | 0.85 | 0.85 | 0.88 | 0.88 | 0.88 |

Худшие результаты продемонстрировал наивный байесовский классификатор на самом простом наборе признаков – частеречных характеристиках, длине слова и длине предложения. Лучше всего с поставленной задачей справляется классификатор случайный лес. Ожидаемым образом, добавление в модель лексических и синтаксических признаков улучшает работу всех трех классификаторов. С помощью функции var\_imp пакета SKlearn мы получили список конструкций каждого глагола, ставших наиболее важными признаками машинного обучения, и рассчитали для них LL-score в каждом подкорпусе, отобрав конструкции с LL-score > 3,5 – маркерные для жанра. Их анализ проводился по следующему алгоритму:

1. Для каждой конструкции определялся семантический класс глагола-вершины.
2. Определялась ядерная (прототипическая) конструкция.
3. Если конструкция, представленная в тексте, отличается от ядерной, можно говорить об усечении или расширении конструкции.
4. Из банка примеров извлекались предложения, соответствующие схеме.
5. Примеры анализировались с опорой на тематику микрожанра.

В **Главе 5** приведены результаты анализа маркерных конструкций. Нами был сформулирован ряд гипотез о причинах сдвигов распределения глагольных конструкций в этих микрожанрах.

1. Жанр произведения представляет различные тематики, что накладывает ограничения на выбор лексики и лексической конструкции. Это проявляется в высокой частоте жанроспецифичных глаголов; кроме того, с тематикой произведения связано отдельное значение (фрейм) лексической единицы, а также выбор его синтаксической конструкции.
2. Авторство также накладывает отпечаток на выбор лексики и конструкций. В нашем исследовании мы постарались минимизировать вклад стиля, ограничив количество произведений одного автора пятью книгами.
3. Разные жанры имеют разные схемы развития сюжета. К примеру, любовным романам присуща линейная структура повествования с ключевыми сценами и сценами-филлерами. Это обуславливает наличие двух групп глаголов с различными синтаксическими конструкциями, при этом лексические конструкции в ключевых сценах воспроизводятся из романа в роман, составляя «формулу», сцены-филлеры не обязаны быть таковыми.

В более широком ключе, можно предположить, что различные жанры имеют разную структуру нарратива, диалога и других литературных форм, что проявляется в наблюдаемых сдвигах распределения лексико-синтаксических единиц.

Мы описали глагольные конструкции, ставшие маркерами литературных формул, используя следующую схему:

* лексико-семантическая группа (группы);
* полная конструкция (ПК), отражающая конвенциональный способ выражения всех обязательных участников семантического фрейма;
* сравнение конструкции-маркера с полной конструкцией с точки зрения:
  + полноты,
  + наличия новых (сирконстантных) участников,
  + особенностей морфосинтаксического оформления участников,
  + особенностей лексического профиля заполнения слотов конструкции.

Маркерные глагольные конструкции (всего 81) были объединены в 13 лексико-семантических групп, из которых наиболее крупными были ментальные глаголы, глаголы движения и речи. Интересно отметить, что лексическими маркерами поджанра могут быть как жанроспецифичные глаголы, так и целые группы (глаголы социальных взаимодействий могут описывать психологическую прозу, глаголы движения – боевики и т.д.).

Полная конструкция (ПК) – конвенциональный способ выражения всех ядерных участников семантического фрейма (т.е., конструкция, реализующая все валентности глагола в данном значении); для определения ПК использовались Малый академический словарь русского языка, ФреймБанк, Активный словарь.

Затем примеры из текста, реализующие маркерную конструкцию, сравнивались с ПК с точки зрения количества аргументов (конструкция в примере полная, неполная, расширенная) и состава зависимых (как реализованы те или иные связи, какая лексика заполняет валентности) и системные расхождения комментировались. Всего нами было подробно разобрана 151 конструкция, маркирующих один или несколько микрожанров. Ниже приведен пример разбора значимой для любовных романов конструкции <*краснеть*/*розоветь* *nsubj*>.

**Маркерная конструкция: nsubj**

    лексико-семантический класс: физиология

        конструкция глаголов *краснеть*, *розоветь*

        ПК: V + nsubj (кто?/что?)

        ПК: V + nsubj (кто?) + obl (от чего?) + [obl (до какого предела?)]

*Лицо продавца* ***порозовело****.*

*Пока она* ***краснела*** *и пятилась, он смотрел на письмо, зажатое в ее руке, и, наконец, протянул свою, явно собираясь отобрать бумагу.*

Жанр любовного романа способствует реализации значения «эмоциональная (физиологическая) реакция» у глаголов изменения цвета. Реакция может быть результатом или преддверием плача, а также связана со смущением или эротическими переживаниями. Конструкция-маркер – субъектная: позиция субъекта заполняется названием лица (краснеют, прежде всего, героини) или наименованием частей тела (упоминаются глаза, личико, щеки, уши, грудь). По сравнению с полной конструкцией, в конструкции-маркере жанра не называется причина, вызвавшая эмоцию. Как правило, эмоция уже упомянута в предшествующем контексте, а синтаксис любовного романа избегает излишнего усложнения.

В результате анализа маркерных глагольных конструкций мы выделили полные (с реализацией всех обязательных валентностей) и неполные (с опущением одной или нескольких из них): бессубъектные (с опущением подлежащего), безобъектные (с опущением прямого объекта), прочие (с опущением других обязательных актантов). Кроме того, были проанализированы расширенные конструкции, в которых, помимо обязательных актантов, были выражены сирконстантные участники. Полные конструкции более характерны для фэнтези и научной фантастики, что может быть связано с большей описательностью этих жанров, требующих более связного текста. Неполные конструкции чаще встречаются в любовных романах и детективах – как бессубъектные, так и безобъектные.

Большая часть бессубъектных конструкций приходится на употребление глагола в составе деепричастного оборота или сочиненной группы. Опущение актора – серьезная модификация глагольной конструкции, которая может говорить о сравнительной сложности синтаксиса: как минимум, в таких предложениях несколько глаголов, усложняющих синтаксическое дерево, и поэтому актор часто синтаксически контролируется другим предикатом. Ср. бессубъектную конструкцию в следующих примерах:

*(LS) Мой партнер, не* ***убирая*** *руки с моей талии, нагнулся к моей кисти и поцеловал.*

*(DS) – Это Рип, – сказал я,* ***убирая*** *руку с рукоятки «зиг-зауэра».*

*(FT) – И кто у нас тут самый глазастый, – буркнула лучница,* ***убирая*** *стрелу в колчан.*

В отличие от субъекта, объект опускается гораздо реже, и соответствующая неполная конструкция практически не наблюдается среди конструкций-маркеров формульной литературы. Единственным исключением стала конструкция глагола *глотать* с инкорпорированным объектом (маркер детектива). Невыражение объекта может объясняться конкуренцией разных способов выражения второго семантического актанта: конструкции с прямым объектом и конструкции с комплементной клаузой, ср. *боялся насмешек* и *боялся показаться смешным / что его дядя-лесоруб посчитает племянника плаксой*. В русском языке крайне редко прямой объект и комплемент относятся к разным семантическим актантам и выражаются одновременно – чаще это два способа выражения одного участника. Например, прямой объект может сжато выражать пропозицию, которую иначе в развернутом виде выражал бы комплемент (*решить вопрос*).

Среди конструкций с невыраженным объектом мы наблюдали следующие группы: конструкции с конкуренцией моделей управления xcomp и obj, вытесняющие прямое дополнение (характерны для фентези, большая часть – модальные и бытийные глаголы); конструкции, где объект вытеснен на периферию (объект существует, но имеет низкий коммуникативный ранг: *рассказывать о происшествии*, но не *рассказывать сказку*, конструкция с предлогом может восприниматься как более нейтральная); конструкции, сопровождающие прямую речь (характерны для любовных романов и детективов, тема разговора вынесена в реплику).

Опущения иных семантических актантов мы рассмотрели, разделив их по типу семантической роли. Первую группу образуют глаголы с невыраженной валентностью начальной / конечной точки / локации – глаголы движения и трансформации, бытийные глаголы. В маркерных конструкциях ожидаемо выражается одна из периферийных валентностей, на которую ставится акцент в самом семантическом фрейме или коммуникации. Далее следуют глаголы с невыраженной валентностью адресата / темы / содержания высказывания – глаголы речи: что семантические актанты «адресат» и «тема» уже выражены в реплике персонажа и не повторяются в комментарии к прямой речи, содержащей целевой глагол. Среди других конструкций-маркеров, в которых не выражен косвенный объект, наблюдаются конструкции с участником способ (для звонить (*по телефону* или *в дверь*)), точка контакта (для *подхватывать* (на руки, *под локоток*)), стимул (*смеяться* (*над кем* или *от чего*), краснеть, розоветь (от *удовольствия*)), содержание пропозиции (подозревать *в чем*). На наш взгляд, причин опущения несколько, и они не образуют системной картины.

Другой тип конструкции, отличающихся от прототипических – «расширенные» конструкции, в которых выражается сирконстантный участник. Связь obl может указывать как на обязательного, так и на сирконстантного участника фрейма, однако по нашим оценкам доля выражения сирконстантов в конструкциях-маркерах, содержащих obl, невелика. Связь advcl почти всегда соотносится с сирконстантом. Можно отметить такие глаголы, как *просить* и *бояться*, у которых advcl кодирует как обязательного участника (содержание, придаточную клаузу с союзами *чтобы* и *если*), так и обстоятельственного (время, цель и т.д.). У трех конструкций участник advcl всегда является сирконстантным: <*бродить nsubj;obl;advcl*> <*подниматься nsubj;obl;advcl*>, <*стать nsubj;xcomp;advcl*>. В формульной литературе данные глаголы с конкретным значением могут служить своеобразными «точками доступа» к описанию эмоциональных переживаний, интерпретации действий героев (что позволяет раскрыть выразительную описательную, эмоциональную или интеллектуальную составляющую без использования абстрактной лексики). Для упрощения создания эмоциональной вовлеченности синтаксически в центре предложения находится физический глагол, а абстрактный глагол помещается в зависимую клаузу.

*Загрузив провизию в машину, я* ***стала*** *придумывать меню на вечер.*

*Наверху Рэн* ***бродила*** *по комнате, заламывая руки и вытирая бегущие слезы.*

*Сердце словно вырвали у нее из груди, волна слез* ***поднималась*** *к глазам, грозя затопить печалью.*

Предложенный метод использования поверхностно-синтаксической информации, безусловно, имеет ряд ограничений. Во-первых, он опирается на автоматическую обработку больших массивов текстовой информации, которая не свободна от ошибок, хотя ситуация неуклонно улучшается.

Во-вторых, метод был опробован на материале четырех формул массовой литературы, и его успешность в отношении других выборок для корпуса, других микрожанров литературы, других разновидностей текста и авторства еще только предстоит оценить.

Третье ограничение связано с тем, что из корпуса извлекается только поверхностная синтаксическая информация, в то время как глубокий синтаксис, семантика и прагматика конструкции, структура нарратива и дискурса малодоступны как признаки автоматической разметки при нынешнем уровне развития корпусных технологий для русского языка. Дальнейшее развитие метода здесь видится в совершенствовании аппарата описания взаимодействия синтаксиса и семантики в художественном тесте и появлении точных алгоритмов автоматического обнаружения соответствующих признаков.

Предложенный метод выделения конструкций не различает случаи, когда у глагола имеется одно vs несколько значений. Многозначные глаголы имеют существенные различия в частоте употребления, но объем нашего исследовательского корпуса недостаточен для полноценного исследования этого явления - однако наш метод дает ключ к извлечению такого материала.

Наше исследование строится в парадигме эмпирической, usage-based лингвистики, в которой лингвистический материал, извлеченный статистическими методами из корпуса, служит базой для формулировки новых лингвистических гипотез и объяснений. Очевидны перспективы развития самих компьютерно-количественных методов — классификации, кластеризации, ранжирования, также с добавлением новых признаков, как явно представленных в контексте, так и сжатых векторных представлений и скрытых категорий — для построения конструктиконов литературного текста. Вместе с тем, сам собранный нами корпус формульной литературы в дальнейшем может быть использован как стандартный набор данных для оценки качества моделирования поведения конструкций в тексте.

Публикации

Буйлова Н. Н. Классификация текстов по жанрам при помощи алгоритмов машинного обучения // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2018. № 8.— С. 34-38.

Буйлова Н. Н., Ляшевская О. Н. Лексико-синтаксические маркеры малых литературных жанров // Вестник Православного Свято-Тихоновского гуманитарного университета. Серия 3: Филология. 2021. Т. III. № 66. — С. 11-23.

Builova N. Verb constructions as a feature of genre classification, in: CLLS 2018. Computational Linguistics and Language Science. Proceedings of the Workshop on Computational Linguistics and Language Science. Moscow, Russia, April 25, 2018 / Ed. by E. L. Chernyak. CEUR Workshop Proceedings, 2018. pp. 108-113.

Byjlova N. Amateur Prose on the Web: Verb Construction as a Feature of Genre Classification, in: Anna Butašová, Vladimír Benko, Zuzana Puchovská [eds.), Proceedings of ARANEA 2018. Web Corpora as a Language Training Tool, 2018. Bratislava: Univerzita Komenskeho v Bratislave. ISBN: 978-80-223-4597-2. p 25-30.

Ахапкина Я. Э., Буйлова Н. Н. Нестандартная русская речь: корпусные технологии в исследовании и методике преподавания // В кн.: Проблемы преподавания курса "Русский язык и культура речи в вузах". М. : Общество с ограниченной ответственностью "Научный консультант", 2016. С. 48-51.

Буйлова Н. Н., Ляшевская О. Н. Так говорят анимешники // В кн.: Современный русский язык в Интернете / Отв. ред.: Е. В. Рахилина, Я. Э. Ахапкина. М. : Языки славянских культур, 2014. С. 239-249.