УДК 004.8

# РОЛЕВАЯ СТРУКТУРА ПРЕДИКАТНЫХ СЛОВ В РЕШЕНИИ ЗАДАЧ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ТЕКСТОВ СОЦИАЛЬНЫХ МЕДИА1[1]

***Никитина Елена Николаевна,***
*кандидат физических наук, научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), Москва, yelenon@mail.ru*

***Смирнов Иван Валентинович,***
*научный сотрудник ГУ ИПИИ, кандидат физико-математических наук, доцент, заведующий отделом «Интеллектуальный анализ информации» ФИЦ ИУ РАН, ivs@isa.ru*

**Аннотация**

В статье дается анализ теории синтаксем Г.А. Золотовой и прикладных исследований, осуществляемых на этой основе. Особое внимание уделяется описанию эмотивного класса лексики в плане частеречной принадлежности, структуры, экспериенциальной и каузативной семантики.

**Ключевые слова:** машинный анализ текста, предикат, синтаксема, существительное, падеж, категориально-семантический класс, семантическая роль, лингвистическая и психологическая интерпретация текста.

## ВВЕДЕНИЕ

Социогуманитарные исследования социальных медиа нуждаются в инструментарии анализа сетевых сообществ и поведения человека в условиях сетевой коммуникации. Такой инструментарий может разрабатываться на основе моделей и методов искусственного интеллекта. Для социогуманитарных исследований большое значение имеет анализ дискурсивного пространства социальных медиа, т.е. анализ текстовой продукции сетевых сообществ, таких как дневниковые записи, реплики в сетевых дискуссиях и т.п. В настоящей статье представлен опыт применения к анализу семантико-синтаксической структуры высказываний в социальных сетях метода реляционно-ситуационного анализа [6, С. 3-10 и др.], который опирается на синтаксемный анализ Золотовой [1] и на концепцию неоднородных семантических сетей Г.С. Осипова [7].

1 Работа поддержана РФФИ, грант №

## Теория синтаксем Г.А. Золотовой

Теория синтаксем профессора Г.А. Золотовой, представленная в «Синтаксическом словаре» [2], занимает особое место среди отечественных систем анализа естественного языка. Большинство из них обращались к слову (обычно глагольному слову) как центральному объекту описания, как это принято в словоцентричных лингвистических концепциях (Н. Хомский, Ч. Филлмор), и прошли путь от решения типологических прикладных задач к становлению и развитию в рамках академической науки (интегральное описание языка Ю.Д. Апресяна, лингвистические труды И.А. Мельчука, изначально направленные на решение задач машинного перевода).

Особое место золотовской теории обусловлено тем, что, во-первых, Г.А. Золотова, будучи ученицей академика В.В. Виноградова, стала преемницей высокой филологической традиции, соединяющей в одном исследовании лексику и грамматику, грамматику и текст; во-вторых (и как следствие первого), теория синтаксем выросла из наблюдений над естественными текстами разной жанровой принадлежности и тематики и была обусловлена исследовательской потребностью понять, как членится текст, предложение, конструкция, на какие естественные минимальные составляющие распадаются неэлементарные синтаксические единицы и объекты. Материалом, с которым работала Золотова, создавая теорию синтаксем, была именная грамматика – имена существительные и замещающие их местоимения. Наиболее полно и плодотворно такой подход был отработан на существительных, далее понятие синтаксемы было распространено на другие части речи (глагол, прилагательное)[2]. Синтаксемами были названы элементарные, далее неделимые единицы синтаксиса, т.е. слова в определенной морфологической форме, отнесенные к определенному категориальному классу, входящие в качестве конструктивного компонента в состав синтаксической конструкции (синтаксической единицы более высокого уровня).

Для грамматической традиции, в которой работала Г.А. Золотова, не характерен механистический расщепляющий уровневый подход к языковым объектам, для виноградовской традиции синтаксис предстает как объединяющий и организующий центр грамматики. Соответственно, синтаксема как единица синтаксиса представляет собой слово в единстве формы (морфологической), значения (категориально-семантического) и функции. Бытие именной синтаксемы предопределяется падежом ***(формой)***, категориально-семантической общностью, в которую включается индивидуальное лексическое значение слова, ***(значением)*** и предназначенностью выполнять определенную роль в рамках конструкции ***(функцией)***.

2 Попытки описания в идеологии теории синтаксем других частей речи в синтаксисе были сделаны и делаются учениками Г.А. Золотовой – М.Ю. Сидоровой (прилагательное), Н.К. Онипенко (глагол), В.Е. Чумирина (глагол).

Эти три аспекта (которые в уровневой лингвистике разведены по разным ярусам языковой системы и, соответственно, изучаются в рамках разных лингвистических дисциплин) неотделимы друг от друга и являются характеристиками синтаксического объекта в реальной конструкции, в реальном тексте. В этом заключается лингвистический реализм Г.А. Золотовой. Именно лингвистическим реализмом объясняется её критика «двухэтажной грамматики» вербоцентрических описаний предложения [2], противопоставляющих семантику как совокупность значимых, смысловых компонентов и синтаксис как лишенный смысловой сущности иерархический или линейный порядок. Для Г.А. Золотовой нет синтаксиса вне семантики. Интересом к имени, его центральной ролью в создании теории синтаксем определяется третье отличие золотовского подхода на фоне вербоцентрических лингвистических построений, в которых любая конструкция (в том числе предложение) рассматривалась как имеющая глагольное ядро (приоритет глагола), а все имена получали статус глагольных зависимых. Золотовская теория показала, что разные именные синтаксемы имеют и разный функциональный потенциал по отношению к синтаксическому контексту (конструкции). Именные синтаксемы образуют трёхчленную функциональную типологию: по способности образовывать значение они делятся на ***свободные, конструктивно-обусловленные, связанные.***

***Свободные синтаксемы*** располагают формой, значением и функцией вне контекста, изолированно (*В поле, на рынке* – локативы; *О любви*, *Про это* – делиберативы, предмет мысли и речи; *В январе, В субботу* – темпоративы) и могут встраиваться в контекст в качестве конституирующих компонентов предложения (субъекта и предиката) либо занимать присловную позицию. ***Обусловленные синтаксемы*** обладают формой, независимой от контекста, а значение приобретают в рамках конструкции, т.е. их значение производно от функции или, другими словами, функция проявляет их значение (*Мама* спит, она устала – носитель психофизического состояния, экспериенцер; *Мне* не спится – носитель психофизического состояния, экспериенцер; *Мама* мыла раму – носитель акционального признака, агенс). ***Связанные синтаксемы*** не располагают ни собственной формой, ни значением, ни функцией на уровне предложения, они занимают присловную позицию, являются продолжением лексического значения предикатного слова, от которого зависит их форма падежа (Мама мыла *раму* – Винительный падеж, беспредложный, объект глагола). Таким образом, в русской синтаксической системе есть как вербоцентрические конструкции (глагольные словосочетания), так и невербоцентрические, с паритетом имени и глагола (предложения с глагольным сказуемым) или приоритетом имени (именные предложения).

В-четвертых, золотовская теория синтаксем была создана в рамках антропоцентрической объяснительной грамматики (предназначенной для понимания человека человеком), применялась как инструмент многоуровневого анализа художественного текста (на ступени элементарных (общих) языковых средств), а затем оказалась востребована в прикладных исследованиях и специализированных лингвистических дисциплинах. Например, синтаксемный анализ положен в основу инструментов машинного анализа текста, разрабатываемых в лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ (О.В. Кукушкина), синтаксема используется в отечественной онтолингвистике как одно из средств интерпретации детской речи и развития речевой способности ребенка.

## Теория синтаксем в исследованиях текстов социальных сетей

Применение теории синтаксем в ряде прикладных междисциплинарных исследований, проводимых в ФИЦ ИУ РАН, показало эффективность и надежность данного инструмента интеллектуального анализа текста [4, 5 и др.]. Сведения о синтаксической сочетаемости каждого глагола с синтаксемами заносятся лингвистами в словарь предикатных слов. Для построения словаря предикатных слов, разрабатываемого в ФИЦ ИУ РАН, было применено схематическое описание именных компонентов синтаксических конструкций (синтаксем) как зависимых глагольного (предикатного) слова. Технически синтаксема в данном описании представляет собой единство трех сторон: 1) падеж, 2) категориальный класс имени, 3) семантическая роль. Отнесённость конкретного имени к определённому категориальному классу уточняет его семантическую роль в конструкции: *работа_pred* ***Игоря/руководителя/музыканта*** (одушевлённые имена – агенс) – *работа_pred* ***двигателя/мотора/лифта/оборудования*** (предметные имена – объект); *письмо написано_pred* ***Игорем*** (агенс) – *письмо написано_pred* ***карандашом*** (инструмент).

Исследования, в которых первоначально применялся синтаксемный анализ, базировались на диалоге человека и машины: оформлялся пользовательский запрос, сформулированный на естественном языке, на поиск определенного содержания в корпусе документов, машина преобразовывала запрос в синтаксемы определенной семантики и осуществляла поиск, пользователь получал результаты поиска на естественном языке. Тем самым синтаксема использовалась в качестве обслуживающего машину элемента.

На следующем этапе междисциплинарных исследований встала новая задача. Специалисты разных областей знания (психологи и лингвисты) должны были получить возможность оценивать семантическую составляющую текста после машинного анализа по семантическим ролям, т.е. по значениям синтаксем, которые теперь выполняют не только техническую роль, обеспечивающую функционирование анализатора, но и являются средством интерпретации текста, его автора и авторских интенций, например при анализе контента социальных сетей. Это повысило требования к точности описания категориальных классов и семантических ролей.

Семантические роли как компонент описания именной грамматики восходят к тем значениям, которые выделены Г.А. Золотовой [2]. Эта книга представляет собой необыкновенно яркий опыт лингвистической рефлексии по поводу синтаксиса имени существительного. Индивидуальный взгляд на синтаксис родного языка и исследовательская независимость привели к свободному употреблению терминологии, роль которой состояла не столько в том, чтобы однозначно зафиксировать некоторое явление, сколько в том, чтобы точнее, тоньше выразить понимание языкового факта. Но исключительный интерес к определённым семантическим типам (модусные синтаксемы, каузативные синтаксемы) предопределил то, что внимание к материалу

было распределено неравномерно, часть синтаксем получила очень подробную интерпретацию, часть – недостаточную. Поэтому при практической опоре на идеи Золотовой при вводе в электронный словарь конкретных предикатных слов разработчикам пришлось упрощать или расширять описание зависимых имен. К примеру, как интерпретировать формы местоимения *Он* в паре коррелирующих высказываний (1) *Он страшится неизвестности*… – (2) *Его страшит неизвестность*…? *Его* в (2) получает двойную характеристику: «объект каузирующего воздействия и субъект каузируемого состояния» [2, 2001 ,С. 157]. Для *Он* в (1) наиболее подходит такая квалификация: «предицируемый субъект статуального признака» [Там же, 23]. Из двух или более семантических признаков техника описания в электронном словаре требует выбрать единственный. Так, для семантического компонента *Его* (2) при выборе единственного признака необходимо семантически упорядочить экспериенциальную роль одушевлённого имени в коррелирующих конструкциях (1) и (2). (См. примеры из диалогов в соцсети Пикабу с корректными разборами анализатора: [Я#экспериенцер] не совсем [интересовался@Предикат] этим вопросом; А [меня #экспериенцер] ещё всегда [интересовал@Предикат] [вопрос #каузатив] насчет ГМО).

Упрощение квалификации именных компонентов конструкций коснулось не только семантической стороны, когда несколько характеристик сводились в одну. Кроме того, на словарном уровне пришлось отказаться от функциональной квалификации, которая для конструктивно-обусловленных синтаксем состоит в своеобразном единстве (или взаимосвязанности) функции и значения – словарное описание ограничивается приписыванием значения. Функциональный аспект может быть при необходимости восстановлен на этапе анализа коммуникативных единиц (предложений). Вычисление субъектного компонента может быть целесообразно для понимания семантического типа его предиката (словарное значение предикатного слова может модифицироваться в производных диатезах), ср.: *Игорь сломал принтер* (личный субъект-агенс, предикат действия) – *Принтер сломан* (= принтер неисправен; предметный субъект-носитель качества, предикат качества) – *Принтер сломался* (= принтер неисправен; предметный субъект-носитель качества, предикат качества). А количественный анализ семантических типов предиката в тексте может дать представление об окраске текста в плане активности/рефлексивности, динамики/статики и т.п. Количественное уменьшение субъектных компонентов относительно предикатных может служить не только для понимания анафорических связей (опущение субъектного компонента не всегда сигнал анафоры), но и для определения внешней или внутренней точки зрения пишущего (модусная составляющая высказывания).

Примеры с опущенным экспериенциальным компонентом, в которых восстанавливается семантика Я:

- Всегда [забавляла@Предикат] странная [убежденность#каузатор] украинцев, что у нас то же самое, что и у них. (= «меня забавляла»);
- Хоть [порадоваться@Предикат] [за человека#каузатив]! (= «я хочу/готов порадоваться»);
- Тоже [этот момент#каузатор] [напряг@Предикат]. (= «напряг меня»).

При этом интересно, что техника вычисления субъектного компонента обратна золотовской логике конструктивно-обусловленной синтаксемы: у Золотовой от функции синтаксемы (предназначенности к выполнению синтаксической роли) к значению (проявляемому в предложении), в прикладном исследовании от значения синтаксемы (описанного в словаре) к её функции (в предложении).

Ещё одна трудность описания была связана со статусом свободной синтаксемы, которая несёт значение в изолированном употреблении и при встраивании в контекст значение сохраняется. На этом основании, в целях экономности, первоначально описание в словаре предикатных слов было сделано таким образом, что все свободные синтаксемы получали разборы без учета предикатного ядра. Однако в русском языке существуют омонимичные пары свободных синтаксем: каузативное ИЗ-ЗА+Род. (*они поссорились* ***из-за стола***) и директивное (пространственное) ИЗ-ЗА+Родительный падеж. (*Он раздражённо отшвырнул газету и вышел* ***из-за стола***); каузативное С+Родительный падеж (*бросил* ***с досады***) – директивное С+ Родительный падеж (*бросил* ***с балкона***). Если человеческое сознание легко разрешает омонимию контекстно, то при машинном анализе возникала ситуация неоднозначности, конкуренции и неправильного выбора машиной семантической роли. См. примеры, где машина некорректно произвела разметку директивов вместо каузативов: (3) «*Корейцев» и Ладу покупают из-за низкой* ***цены*** *и страха разбить первый автомобиль* (сайт pokupka_avto); (4) *с* ***чего*** *это плебесцит не имеет юридической силы, в какой стране мира?* Если в (3) омонимия могла бы разрешаться корректным выбором категориального класса: имя-характеристика (*цена*) не должно получать разбор посредством пространственной семантической роли, то ошибка в (4) преодолевается только признанием того факта, что разные свободные синтаксемы имеют разный ранг. Причинность (реализуемая каузативами) встраивается практически в любой контекст, т.к. практически любое положение дел мы можем осмыслить в категориях причинно-следственных отношений. Пространственная динамика (реализуемая директивами) требует соответствующего семантического компонента от предикатного слова – на основаниях «изотопии», или «семантического согласования» (Ж. Женетт, Ю.Д. Апресян), поэтому директивные синтаксемы должны получать описание только при глаголах, чья семантика предполагает пространственную динамику, ср. (5) с директивом и (6) с каузативом: (5) *Сначала мы с мужем пробовали ставить рядом с ним будильник и потом, через час, когда он звенел, выгонять Павлика* ***из-за компьютера***. – (6) *Ну а потом стало ясно, что главная проблема — это не глаза и не шумные компании, а то, что* ***из-за*** *своего* ***компьютера*** *Павлик совершенно забросил школу и перестал заниматься*[3].

Некоторые новации в репертуар семантических ролей были внесены по запросу психологов, интерпретирующих сетевую активность с точки зрения деструктивности, агрессивности настроений пользователей. Так,

3 Примеры из Национального корпуса русского языка (НКРЯ).

для этих целей в рамках семантической роли объекта было важно разграничить собственно объект (нейтральный): мама мыла *раму*, и два негативных: деструктив: мама поцарапала раму, ликвидатив: мама сломала *раму*. В этом заключается определенное отступление от собственно лингвистической описательной идеологии, которое может быть преодолено за счет более тщательной квалификации класса глагола (а не объекта), так как деструктивность семантики принадлежит именно глаголу, а объект, будучи связанной синтаксемой, значения не несет.

## Приемы и методы пополнения словаря предикатных слов

В настоящее время в рамках расширения электронного словаря предикатных слов реализуется проект комплексного описания слов одной категориальной семантики на единых структурно-семантических основаниях, что позволяет применять компактное единообразное семантическое описание на большом массиве лексики. В качестве материала используется микрословарь эмотивных каузативов и их возвратных коррелятов, а также предикативных наречий (категория состояния) и существительных, сохраняющих то же эмотивное значение и способных выступать в качестве предикативного ядра (*возмущать – возмущаться – возмущение – возмутительно; страшить – страшиться – страх – страшно*), всего около 300 слов. Семантическая (ролевая) модель данных предикатных слов содержит компоненты («семантические роли») экспериенцера (личного субъекта состояния: *его возмущает..., он возмущается..., ему возмутительно..., его возмущение...*) и – в разной мере – каузатора (причина состояния: *возмущает поведение; возмутительно, что..., он возмущается, что...*).

В ролевой структуре возвратного эмотивного глагола сема причины вытесняется (что отражается в морфемной структуре глагола: позиция каузатора закрывается постфиксом). Однако эмотивная сфера такова, что человек испытывает потребность мотивировать эмоциональные состояния словом, поэтому семантика причины находит выражение в подобных высказываниях с помощью разнообразных «периферических» каузативных синтаксем. Для описания причинной семантики при возвратных глаголах и коррелирующих существительных использовался широкий подход: помимо закрепленных в литературной норме были обработаны нелитературные формы, т.к. они широко представлены в сетевом контенте: не только *возмущение этим фактом*, но и *Свое возмущение этому* ***поводу*** *уже выразили проживающие в Великобритании мусульмане; возмущение* ***на*** *чужие* ***грехи;*** Так вы же сами [возмущаетесь@Предикат] [отпискам#каузатор](Пикабу).

Подобный подход применен и к словам на *–о* – дериватам эмотивов, многие из которых соединяют в себе свойства двух категорий – состояния и оценки, находясь на границе двух этих классов слов, ср.: ***Мне*** *страшно* (категория состояния; литературная норма) – *Если коротко, то* ***мне*** *возмутительно, что крупный оператор ведет себя как мелкий обманщик* (отзыв на banki.ru). Категория состояния как тип предиката предполагает индивидного экспериенциального субъекта в датательном падеже, в то время как оценка выражает общее, коллективное мнение, «как надо», позиция для неиндивидного субъекта-носителя всеобщего мнения в предложении отсутствует (*Курить вредно*).

В плане прикладном данный семантический класс и широкий подход к описанию падежных форм, которыми выражается каузативная семантика, позволит составить

представление о том, кто, по мнению пользователей соцсетей, является субъектом называемых состояний, «измерять» настроение и эмоциональные состояния участников сетевых дискуссий и судить о причинах испытываемых и выражаемых словесно состояний.

Если на машинном анализе контента соцсетей обсуждаемый лингвистический материал – в аспекте анализа ролевой структуры глагола – только предстоит опробовать, то в области изучения лингвистических особенностей текстов с точки зрения психиатрической нормы и патологии уже получены первые научные результаты: данный семантический класс применялся для выявления значимых различий в употреблении эмотивной лексики разной структуры (возвратные/невозвратные глаголы), формы (Я/не-Я экспериенцеры, Я/не-Я-каузаторы) и лексической семантики (негативной, позитивной, амбивалентной и деэмотивной) в текстах лиц разного психиатрического статуса (депрессия и шизофрения)[3 (в печати)].

## ЗАКЛЮЧЕНИЕ

Практика применения теории синтаксем Г.А. Золотовой в прикладных описаниях для целей интеллектуального анализа текста свидетельствует, что надёжные и непротиворечивые результаты для лингвистов и психологов, интерпретирующих человека посредством его речевой продукции, могут быть получены только при корректном переложении лингвистических идей с языка академической науки на язык информационных технологий. Для этого следует признать необходимость словаря как технологически необходимого, но не единственного этапа анализа. Словарь – это проявление атомарного подхода к (предикатной) лексеме, хранящей семантику, разворачиваемую в предложение. Далее следует проекция словаря на текст, на этом этапе неизбежно должен осуществляться синтез семантической и функциональной составляющих, чем обеспечивается бытие предложения в тексте (и понимание предложения: его типа субъекта, типа предиката, синтаксических нулей и их роли в предложении и тексте). А это значит, что в прикладных исследованиях текста будут востребованы и теория синтаксем (элементарных синтаксических единиц) и другие идеи Золотовой, затрагивающие синтаксис предложения и текста и отражённые в её концепции коммуникативной грамматики, которую на современном этапе продолжают развивать её ученики.

Создание средств мониторинга социальных сетей и моделей протекания массовых психических процессов, на основе описанных в этой статье принципов, позволит проектировать системы прогноза реакций населения в интересах повышения качества управленческий решений локального, регионального и федерального уровней на основе методов искусственного интеллекта.

## Литература


1. *Золотова Г.А.* Коммуникативные аспекты русского синтаксиса. — М., 1982.
2. *Золотова Г.А.* Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. — М., (1988) 2001.
3. *Крылов С.А., Никитина Е.Н., Онипенко Н.К., Станкевич М.А.* Типы грамматической информации в «Семантико-грамматическом словаре русских глаголов» и их применение в машинном анализе текста (на материале глаголов с каузативно-эмотивной семантикой) // Международная конференция «Лингвистический форум 2020: Язык и искусственный интеллект». 12–14 ноября 2020 г. Институт языкознания РАН, Москва: Тезисы докладов (в печати).
4. *Ениколопов С.Н., Кузнецова Ю.М., Смирнов И.В., Станкевич, М.А., Чудова, Н.В.* Создание инструмента автоматического анализа текста в интересах социо-гуманитарных исследований. Часть 1. Методические и методологические аспекты // Искусственный интеллект и принятие решений. — 2019. — №. 2. — С. 28-38.
5. *Кузнецова Ю.М., Смирнов И.В., Станкевич М.А., Чудова Н.В.* (2019).Создание инструмента автоматического анализа текста в интересах социо-гуманитарных исследований. Часть 2. Машина РСА и опыт ее использования // Искусственный интеллект и принятие решений. — 2019. — №. 3. — С. 40-51.
6. *Осипов Г.С., Смирнов И.В., Тихомиров И.А.* Реляционно-ситуационный метод поиска и анализа текстов и его приложения // Искусственный интеллект и принятие решений. — №2 М: ЛЕНАНД — 2008. — С. 3-10.
7. *Осипов Г.С.* Приобретение знаний интеллектуальными системами. — М.: Наука. Физматлит, 1997.
8. *Shelmanov A.O., Smirnov I.V.*, Methods for Semantic Role Labeling of Russian Texts // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue» (2014). Issue 13 (20). — 2014. — pp. 580–592.

# PREDICATE-ARGUMENT STRUCTURE FOR INTELLIGENT TEXT ANALYSIS OF SOCIAL MEDIA CONTENT

***E.N. Nikitina,***
*Candidate of Physical Sciences, Researcher at the Federal Research Center «Informatics and Management» of the Russian Academy of Sciences (FIC IU RAS), Moscow, yelenon@mail.ru*

***I.V. Smirnov,***
*Candidate of Physical and Mathematical Sciences, Associate Professor, Head of the Department «Intellectual Analysis of Information» of the Russian Academy of Sciences, ivs@isa.ru*

**Abstract**

The article provides analysis of the syntactic theory (theory of elementary syntactic units) by G.A. Zolotova and its applications to automated text analysis. The main attention is paid to ways of description of emotive verbs and their derivatives in such aspects as part of speech, structure and semantics of experiencer and cause.

**Keywords:** automated text analysis, predicate, minimal syntactic unit, noun, case, semantic class of lexical unit, semantic role, linguistic and psychological text interpretation