

Научная статья

УДК 81'33+811.161.1+612

## **НЕОДНОЗНАЧНОСТЬ И СИНКРЕТИЗМ КОНЦЕПТУАЛЬНОЙ СТРУКТУРЫ РУССКОЯЗЫЧНОГО КОРПУСА ПРЕДМЕТНОЙ ОБЛАСТИ**

**Анастасия Юрьевна Зиновьева**

Южно-Уральский государственный университет, Челябинск, Россия, zinovevaai@susu.ru, ORCID 0000-0002-7658-7376

**Аннотация.** Статья посвящена явлениям концептуальной неоднозначности и концептуального синкретизма в русскоязычном корпусе предметной области «Исследование интегративной физиологии спортсменов», которая представляет собой актуальное сочетание различных областей медицины и спорта. Изучение этих явлений важно для создания концептуально размеченных корпусов, что имеет большое значение для разработки ресурсов автоматической обработки текста. В статье приведены статистические характеристики распределения концептуально неоднозначных и синкретичных лексических единиц в корпусе указанной предметной области, выявлены источники концептуальной неоднозначности и синкретизма, определены их типы и композиционный состав единиц, характеризующихся этими явлениями. Результаты вносят вклад в понимание концептуальной структуры дискурса предметной области «Исследование интегративной физиологии спортсменов» и могут быть использованы для концептуальной разметки корпуса данной предметной области, а также при разработке правил и метрик снятия концептуальной неоднозначности.

**Ключевые слова:** исследование интегративной физиологии спортсменов, концептуальная неоднозначность, концептуальная разметка, концептуальный синкретизм, предметная область, русский язык

**Для цитирования:** Зиновьева А. Ю. Неоднозначность и синкретизм концептуальной структуры русскоязычного корпуса предметной области // Вестник Челябинского государственного университета. 2023. № 2 (472). Филологические науки. Вып. 131. С. 38–46.

Original article

## **CONCEPTUAL AMBIGUITY AND SYNCRETISM IN A RUSSIAN DOMAIN CORPUS**

**Anastasiia Yu. Zinoveva**

South Ural State University, Chelyabinsk, Russia, zinovevaai@susu.ru, ORCID 0000-0002-7658-7376

**Abstract.** This paper focuses on conceptual ambiguity and syncretism in a Russian corpus on Integrative Physiology, a topical blend of various aspects of medicine and sport. These phenomena need to be studied to produce conceptually annotated corpora, which in turn is of great importance for developing advanced resources for natural language processing. In this article, we provide statistical characteristics of conceptually ambiguous and syncretic lexical units in the corpus on Integrative Physiology, reveal the sources of conceptual ambiguity and syncretism, determine types thereof and composition of the units characterized by these phenomena. The results obtained contribute to the understanding of the conceptual structure of the Integrative Physiology domain and can be used both for conceptual annotation of the domain corpus and for the development of disambiguation rules and metrics.

**Keywords:** conceptual ambiguity, conceptual annotation, conceptual syncretism, domain, integrative physiology, Russian language

**For citation:** Zinoveva AYu. Conceptual ambiguity and syncretism in a Russian domain corpus. *Bulletin of Chelyabinsk State University. 2023;(2(472):38-46. (In Russ.)*

## Введение

В современной корпусной лингвистике глубокий интерес ученых вызывает семантическая разметка, то есть процедура и результат обогащения контента различной семантической информацией. Подобные исследования имеют большое значение как с теоретической точки зрения, позволяя изучить семантическую структуру дискурса, так и с прикладной, обеспечивая основу для разнообразных исследований по автоматической обработке текста. При этом понимание семантической разметки варьируется: среди возможных интерпретаций выделяют выбор значения многозначных слов на основе словаря или онтологии [9], определение семантических свойств слов на основе лексической классификации [3], выявление семантических отношений в тексте [10]. Частный случай семантической разметки представляет собой концептуальная разметка, основанная на какой-либо концептуальной модели предметной области [6; 13]. Процесс концептуальной разметки может быть осложнен концептуальной неоднозначностью лексических единиц [6; 13].

Проблема концептуальной неоднозначности в текстах предметных областей исследована недостаточно. Среди немногочисленных примеров работ по данной тематике можно отметить статьи [13], где описан алгоритм машинного обучения для снятия концептуальной неоднозначности и [6] где упомянута проблема неоднозначности концептуальных тегов в предметной области «Информационная безопасность».

Ранее автором статьи в соавторстве с сотрудниками НОЦ «ЛИНТ» ЮУрГУ (НИУ) было проведено исследование концептуальной неоднозначности при разметке русскоязычных новостных сообщений предметной области «Терроризм», где были рассмотрены источники концептуальной неоднозначности в указанной предметной области и возможные способы снятия неоднозначности с помощью количественных методов [2]. Помимо концептуальной неоднозначности в работе было обнаружено явление концептуального синкретизма, при котором лексическая единица имеет два и более не противоречащих друг другу концептуальных значения, которые не требуют разрешения, однако могут влиять на процедуру снятия концептуальной неоднозначности.

Отметим, что каждая предметная область имеет свои особенности, в том числе на концептуальном уровне, которые следует анализировать для успешного создания концептуально аннотированных ресурсов.

Целью настоящей работы является изучение проблемы концептуальной неоднозначности и концептуального синкретизма в текстах предметной области «Исследование интегративной физиологии спортсменов» (далее — ПО ИИФС) для последующего использования полученных данных в концептуальной разметке текстов данной предметной области. Интегративная физиология, как отмечают исследователи, это «важнейшая для медицины область физиологии, ее вершина» [7. С. 3], которая востребована обществом, системой здравоохранения и медициной. Кроме того, интегративная физиология спортсмена сочетает в себе разнообразные аспекты жизнедеятельности спортсмена от видов тренировок до взаимодействий химических веществ на клеточном уровне, следовательно, концептуальная структура данной ПО разнородна и представляет особый интерес в рамках концептуальных исследований.

Статья организована следующим образом: сначала описаны материалы и методы исследования, затем приведены количественные и качественные результаты исследования неоднозначности и синкретизма концептуальной структуры русскоязычного корпуса ПО ИИФС. В заключении сделаны выводы о проведенной работе и намечены перспективы дальнейших исследований.

## Материалы и методы исследования

Исследование проведено на материале русскоязычного корпуса научных статей ПО ИИФС, опубликованных учеными Южно-Уральского государственного университета в журналах по медицине, физиологии и спорту, в частности «Человек. Спорт. Медицина» (ранее «Вестник ЮУрГУ. Серия „Образование, здравоохранение, физическая культура“»), «Теория и практика физической культуры» и т. п. объемом 54 444 словопотребления. В целях исследования корпус был автоматически размечен концептуальными тегами предметной области с использованием платформы концептуального аннотирования ПАНТ [8], которая была адаптирована для проведения настоящего исследования следующим образом.

Сбор знаний для внесения в программную оболочку ПАНТ выполнялся с помощью прескриптивно-дескриптивной методики. На прескриптивном этапе, исходя из данных лексикографических источников по исследуемой предметной области, был задан набор первичных концептуальных классов (под концептуальными классами в работе понимаются семантические классы,

содержащие лексические единицы конкретной предметной области). Затем из корпуса с помощью экстрактора LanaKey [12] были извлечены релевантные лексические единицы длиной от одного до четырех компонентов и распределены по заданным концептуальным классам. При этом для релевантных единиц, не вошедших по результатам распределения ни в один из заданных классов, были сформированы новые классы, а концептуальные классы, оставшиеся незаполненными, были удалены из набора, в чем заключается дескриптивный компонент методики. В результате был получен список классов, отражающий концептуальную структуру ПО ИИФС в русскоязычном корпусе.

После адаптации платформы ПАНТ к предметной области корпус текстов был размечен автоматически, а затем полуавтоматически доведен до «золотого стандарта»<sup>1</sup> путем снятия неоднозначности с помощью встроенного постредактора. Следует отметить, что автоматическая разметка корпуса формализует информацию о свойствах лексических единиц в предметной области, отражая их возможную концептуальную неоднозначность и синкретизм, в то время как «золотая» разметка отражает контекстную реализацию концептуальных значений лексических единиц и дает контекстно обусловленную информацию о том, в каких случаях требуется обработка концептуальной неоднозначности и концептуального синкретизма для удаления лишних тегов. В связи с этим явления неоднозначности и синкретизма концептуальной структуры были изучены посредством сравнительно-сопоставительного анализа автоматического и «золотого» корпуса.

### Концептуальные классы ПО ИИФС

Общее количество выделенных концептуальных классов составило 39. В табл. 1 представлены десять концептуальных классов с наиболее высокой частотой встречаемости в исследуемом корпусе ПО ИИФС с присвоенными ими тегами для концептуальной разметки и примерами лексических единиц.

### Статистические результаты исследования

Общее количество размеченных лексических единиц (одно- и многокомпонентных) в корпусе

<sup>1</sup> Термины «золотая» разметка, «золотой стандарт», «золотой» корпус обозначают разметку корпуса, выполненную человеком, и корпус текстов с такой разметкой соответственно, см. [11].

с неснятой концептуальной неоднозначностью составило 40 084, из них концептуальных мультитегов — 6 165, из которых 3 486 синкретичны, 2 679 неоднозначны. Наиболее частотные неоднозначные и синкретичные мультитеги приведены в табл. 2 (символом «\*» отмечены мультитеги, представленные как среди неоднозначных, так и среди синкретичных мультитегов).

Результаты анализа мультитегов с учетом лексики показывают, что многие высокочастотные неоднозначные мультитеги закреплены за конкретными лексическими единицами: тег MU-TM («Единицы измерения величин» / «Медикаменты») в исследуемом корпусе во всех случаях присвоен лексической единице *c*, тег AA-W («Возрастные характеристики ЛЗС» / «Момент или период времени») — единице *год*, AG-MU («Половая принадлежность ЛЗС» / «Единицы измерения величин») — единице *м* и т. п.

Полученные данные свидетельствуют в целом о невысоком уровне концептуальной неоднозначности в текстах ПО ИИФС; тем не менее такая проблема присутствует и может вызвать затруднения при последующей обработке размеченного текста, например при извлечении релевантной информации.

### Источники концептуальной неоднозначности

Лексическая единица считается концептуально неоднозначной, если в зависимости от контекста она может иметь противоречащие друг другу концептуальные значения. В корпусе ПО ИИФС концептуальная неоднозначность в основном наблюдается среди однокомпонентных единиц, может встречаться у двух- и трехкомпонентных единиц, имеющих в своем составе родовое слово (например, *гормон кортизол*) или представляющих собой названия химических веществ (например, *йодид калия*). Выявлено несколько источников концептуальной неоднозначности: частеречная омонимия, лексическая неоднозначность, множественность концептуальных значений.

**Частеречная омонимия.** Данный источник имеет место в случае совпадения по крайней мере одной словоформы у разных лексем. Например, словоформа *легкие* может быть существительным, обозначающим орган (в этом случае она относится к концептуальному классу «Локализация биохимических и механических процессов в организме ЛЗС»), или прилагательным, обозначающим весовую категорию спортсмена («Спортивная квалификация ЛЗС»).

Таблица 1  
Table 1

**Наиболее частотные концептуальные классы ПО ИИФС**  
**The most frequent conceptual classes of the Integrative Physiology domain**

Тег	Концептуальный класс	Лексические примеры
OL	Локализация биохимических и механических процессов в организме лиц, занимающихся спортом (ЛЗС)	внутриклеточный процесс, гипофиз, клетка, надпочечник, объем межклеточного пространства, печень, рецептор, соединительно-тканная структура, ткань
MP	Любые измеряемые величины	вариабельность показателей, динамика частоты дыхания, дыхательный объем, содержание липопротеидов высокой плотности
OB	Биохимические процессы в организме ЛЗС	белковый синтез, выделение оксида азота, гидролиз гликогена, гомеостаз, закисление, мышечный метаболизм, скорость продукции CO <sub>2</sub> , электролитный и водный обмен
J	Общая методология и представление результатов научных исследований	автор, доказательная база, исследование, контрольная группа, материал, обсуждение, результаты ранее опубликованных исследований, рисунок, эксперимент
KR	Типы и процессы тренировок	воздействие физических упражнений на выносливость, околопредельный подъем груза, состояние снижения гравитационной и осевой нагрузки, стретчинг
MU	Единицы измерения величин	балл, градус, Дж, дин, ед., ккал/ч, км, л, м <sup>2</sup> , мин, мкг, мл, ммол, процент, с, см <sup>3</sup> , сотая доля секунды, субъединица, уд., ч, час, шт.
OD	Продукты жизнедеятельности организма ЛЗС	аминокислота, белок плазмы, внеклеточная вода, гормон щитовидной железы, кислый желудочный сок, клеточная вода, молочная кислота, серотонин, холестерин
A	Спортсмены и прочие ЛЗС	бегунья, девушка-подросток, лыжница, пловец-дельфинист, ребенок младшего возраста, спортсмен, человек, хорошо тренированный спортсмен, юный пловец
AP	Физические характеристики ЛЗС	критерий подготовленности бегунов, локально-региональная мышечная выносливость, мышечная работоспособность, специальная выносливость организма
ER	Результаты исследования физиологического состояния ЛЗС	без изменений, варьировать, высокое значение, выявить, достоверно не отличаться, наблюдаться, низкое значение

Таблица 2  
Table 2

**Наиболее частотные мультитеги в корпусе объемом 54 444 словоупотреблений**  
**The most frequent multitaggs in 54,444 word usage**

Наиболее частотные неоднозначные мультитеги		Наиболее частотные синкретичные мультитеги	
Мультитег	Частота	Мультитег	Частота
MU-TM	425	MP-OL	291
AA-W	100	A-S	214
AG-MU	98	OB-OD	135
MU-OL	97	AA-A-AG	131
MP-OD*	92	OL-OP	98
MU-OB	82	OB-OL	95
AP-MP	65	MP-OD*	90
ER-OR	65	AP-OL	89
KR-S	63	MP-OP	82
MU-OP-TM	58	G-MP	76

Еще одним примером является словоформа *животного*, которая может быть интерпретирована как существительное или прилагательное. В первом случае она относится к концептуальному классу «Общая методология и представление результатов научных исследований», поскольку обозначает лабораторных животных, во втором случае — к концептуальному классу «Пища, принимаемая ЛЗС», например: *пищевые добавки растительного, животного и минерального происхождения*.

В свою очередь словоформа *железа* может быть отнесена к концептуальным классам «Общая методология и представление результатов научных исследований» и «Локализация биохимических и механических процессов в организме ЛЗС» (ср. *дефицит железа и магния и щитовидная железа*).

**Лексическая неоднозначность.** В статье под лексической неоднозначностью понимается возможность двух и более интерпретаций лексической единицы в конкретном контексте, вызываемая полисемией или омонимией [5]. Лексическая неоднозначность довольно широко распространена в корпусе ПО ИИФС среди однокомпонентных лексем. Приведем несколько примеров.

Лексема *год* реализует в корпусе два из своих значений: 1. Промежуток времени, равный периоду обращения Земли вокруг Солнца — 12 месяцам (концептуальный класс «Момент или период времени»). 2. Возраст («Возрастные характеристики ЛЗС») [4]. Первое значение реализуется в следующем контексте: *через год занятий в группе обследования факторные веса распределились следующим образом*; второе — в контексте: *в возрасте 18–21 года*.

Далее, лексема *введение* также реализует в корпусе два из своих значений: 1. Процесс действия по значению глагола ввести («Физиологическое воздействие на ЛЗС с целью повышения результативности»). 2. Вступление («Общая методология и представление результатов научных исследований») [1]. Первое значение реализуется в следующем контексте: *введение было проведено соответственно в течение 1, 2, 3, 4 и 5-го дней*; во втором значении данная лексема используется в качестве подзаголовка научной статьи.

Еще одним примером является лексема *лицо*, реализующая в корпусе следующие значения: 1. Передняя часть головы человека («Части тела ЛЗС»). 2. Человек, личность («Спортсмены и прочие ЛЗС») [4]. Первое значение реализуется в следующем контексте: *появление волос на лице, сни-*

*жение тембра голоса, покраснение кожи и прекращение менструаций*; второе — в контексте: *наблюдался чаще у лиц с повышенной толщиной миокарда*.

Интересный пример представляют сокращения *мм, м, мк*. Так, сокращение *мм* может быть расшифровано следующими способами: 1. Мышечная масса («Локализация биохимических и механических процессов в организме ЛЗС»). 2. Миллиметр («Единицы измерения величин»).

**Множественность концептуальных значений.** Под термином «множественность концептуальных значений», или, иначе, «собственно концептуальная неоднозначность» мы понимаем принадлежность лексической единицы к некоторому множеству концептуальных классов при единстве лексического значения, при этом в контексте такая единица может быть отнесена только к одному из этих концептуальных классов.

Например, лексема *занятие* может быть отнесена к одному из трех классов:

1. «*Занятия проводили на тренажерах блочно-го типа в состоянии снижения гравитационной и осевой нагрузки*»<sup>1</sup> («Типы и процессы тренировок»).

2. «*Показатели, характеризующие нервно-мышечные связи, также свидетельствуют о том, что занятия, проведенные во второй группе мужчин, достоверно улучшили моторную память*»<sup>2</sup> («Физиологическое воздействие на ЛЗС с целью повышения результативности»).

3. «*Формула спортизации в школах Татарстана выглядит следующим образом: 2 ч отводится на занятия по выполнению требований Госстандарта плюс 4 ч — вариативная часть*»<sup>3</sup> («Образовательный процесс, включающий общие и специализированные предметы»).

<sup>1</sup> Батуева А. Э., Якушева А. Н. Исследование биоэлектрической активности и пространственно-временных характеристик головного мозга в группах мужчин молодого и зрелого возраста с синдромом дорсопатии на поясничном уровне до и после коррекционного воздействия // Человек. Спорт. Медицина. 2013. Т. 13 (4). С. 94–99.

<sup>2</sup> Батуева А. Э., Якушева А. Н., Эрлих В. В. Возможности занятий на блоковых тренажерах при коррекции постуральных нарушений у молодых мужчин с синдромом дорсопатии // Теория и практика физической культуры. 2014. № 10. С. 50–52.

<sup>3</sup> Черепов Е. А., Ненашева А. В. Современное состояние системы физического воспитания в России: основные проблемы и пути совершенствования // Человек. Спорт. Медицина. 2014. Т. 14 (3). С. 5–18.

Множественностью концептуальных значений может обладать в том числе глагольная лексика, как, например, глагол *увеличиваться*:

1. «Значения фибриногена существенно снижались от лета к осени, затем достоверно **увеличивались** зимой и несколько уменьшались летом»<sup>1</sup> («Результаты исследования физиологического состояния ЛЗС»).

2. «Макрофаги начинают жизнь как моноциты с низкой способностью в борьбе с их инфекционными агентами, разбухая, **увеличиваются** в диаметре в 5 раз»<sup>2</sup> («Механические процессы в организме ЛЗС»).

### Источники концептуального синкретизма

Напомним, что лексическая единица считается концептуально синкретичной, если она одновременно реализует два и более концептуального значения, которые не противоречат друг другу, и, таким образом, относится к нескольким концептуальным классам. Концептуальный синкретизм, как и неоднозначность, вызывает появление мультитегов при концептуальной разметке.

Наблюдаемый в ПО ИИФС концептуальный синкретизм может быть разделен на два типа: аксиоматичный и гипотетичный.

Под **аксиоматичным синкретизмом** понимается такой синкретизм, при котором разрешение мультитега не требуется ни в какой ситуации. Примером аксиоматично синкретичного мультитега является АG-A-S («Половая принадлежность ЛЗС» / «Спортсмены и прочие ЛЗС» / «Виды спорта и спортивной деятельности»), которым размечаются лексические единицы, обозначающие спортсменов в различных видах спорта: *пловчиха*, *фигуристка* и др. В данном случае семантические компоненты, связанные с концептуальными классами, обозначаемыми тегами АG и S, выступают в качестве атрибутов (половая принадлежность и вид спорта) основного семантического компонента лексемы, связанного с концептуальным классом, обозначаемым тегом А.

<sup>1</sup> Эрлих В. В. Сезонные биоритмы системы крови, обменных процессов, функций желудка и печеночных проб у бегунов на средние дистанции // Человек. Спорт. Медицина. 2013. Т. 13 (2). С. 36–41.

<sup>2</sup> Эрлих В. В., Исаев А. П., Корольков В. В., Потапова Т. В. Система крови, гомеостаза, метаболизма и функциональные показатели желудка и печени у легкоатлетов-бегуний на средние дистанции после двадцати дней акклиматизации в верхнем среднегорье // Человек. Спорт. Медицина. 2013. Т. 13 (1). С. 17–21.

Под **гипотетичным синкретизмом**, напротив, понимается такой синкретизм, при котором разрешение мультитега может требоваться или не требоваться в зависимости от контекста. Ярким примером гипотетичного синкретизма является мультитег МР-ОD («Любые измеряемые величины» / «Продукты жизнедеятельности организма ЛЗС»), приведенный в табл. 2, которым размечаются названия веществ, образующихся в организме ЛЗС (например, *липиды*, *энзимы*, *глюкоза* и др.). Как правило, этот тег не требует снятия неоднозначности, когда в значении лексемы подразумевается компонент 'содержание'. Например, в предложении «*Энзимы AST и ALT коррелировали соответственно с концентрацией глюкозы*»<sup>3</sup> лексема *энзимы* должна быть размечена тегами МР и ОD, так как содержание энзимов — измеряемая величина, в предложении «*Энзимы играют большую роль в обменных процессах организма*»<sup>4</sup> — только тегом ОD, поскольку измеряемая величина отсутствует. Таким образом, в случае гипотетичного синкретизма одна из сем представлена имплицитно и наблюдается только в определенном языковом окружении.

Выявлено два источника концептуального синкретизма в ПО ИИФС:

1. Однокомпонентная единица содержит несколько релевантных для предметной области семантических компонентов. Например, единицы *ребенок*, *подросток* в текстах предметной области реализуют значения концептуальных классов «Спортсмены и прочие ЛЗС», «Возрастные характеристики ЛЗС».

2. Каждый компонент многокомпонентной лексической единицы имеет собственное концептуальное значение. Например, единица *мониторинг состояния здоровья* состоит из компонентов *мониторинг* (концептуальный класс «Методы исследования физиологического состояния ЛЗС») и *состояние здоровья* («Физические характеристики ЛЗС»). Этот источник обеспечивает большое количество многокомпонентных синкретичных единиц в корпусе.

<sup>3</sup> Романов Ю. Н., Исаев А. П. Вектор и теснота корреляционных пляд состояний кикбоксеров в процессе интегральной подготовки на заключительных этапах готовности к соревнованиям // Человек. Спорт. Медицина. 2013. Т. 13 (2). С. 133–139.

<sup>4</sup> Эрлих В. В. Сезонные биоритмы системы крови, обменных процессов, функций желудка и печеночных проб у бегунов на средние дистанции // Человек. Спорт. Медицина. 2013. Т. 13 (2). С. 36–41.

Вышеуказанные источники концептуального синкретизма могут также сочетаться в пределах одной лексической единицы: например, четырехкомпонентная именная группа *анализ мочи спортсменов-юношей бегунов* отнесена к шести концептуальным классам: AA-A-AG-EM-OD-S («Возрастные характеристики ЛЗС» / «Спортсмены и прочие ЛЗС» / «Половая принадлежность ЛЗС» / «Методы исследования физиологического состояния ЛЗС» / «Продукты жизнедеятельности организма ЛЗС» / «Виды спорта и спортивной деятельности»), где теги AA-A-AG относятся к компоненту *спортсменов-юношей*, EM — *анализ*, OD — *мочи*, S — *бегунов*.

Корреляции между источниками синкретизма и его типами не выявлено.

### Заключение

В настоящей работе представлены результаты исследования корпуса научных текстов предметной области «Исследование интегративной физиологии спортсменов» на предмет концептуальной неоднозначности и концептуального синкретизма лексических единиц. Полученные результаты дают представление об организации дискурса предметной области на концептуаль-

ном уровне и могут быть полезны в прикладных лингвистических исследованиях по концептуальной разметке текстовых корпусов, которая в свою очередь является основой многих направлений автоматической обработки текста.

В ходе исследования выявлены источники концептуальной неоднозначности и концептуального синкретизма в корпусе ПО ИИФС, определены их типы и композиционный состав единиц, характеризующихся данными явлениями.

Следует отметить, что в ряде случаев может быть затруднительно однозначно провести черту между проявлениями концептуальной неоднозначности и концептуального синкретизма, что обусловлено явлением гипотетичного синкретизма, которое в дальнейшем требует более глубокого изучения.

Полученные данные могут быть использованы при концептуальной разметке корпуса ПО ИИФС, при разработке онтологии исследуемой предметной области (в частности, аксиоматично синкретичные мультитеги могут быть полезны при идентификации атрибутов онтологических концептов), а также при создании правил и метрик снятия концептуальной неоднозначности.

### Список источников

1. Ефремова Т. Ф. Новый словарь русского языка. Толково-словообразовательный. М. : Русский язык, 2001. 2354 с.
2. Зиновьева А. Ю., Шереметьева С. О., Неручева Е. Д. Анализ неоднозначности концептуальной разметки русскоязычного текста // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. 2020. Т. 6, № 3 (23). С. 38–60.
3. Рахилина Е. В., Кобрицов Б. П., Кустова Г. И., Ляшевская О. Н., Шеманаева О. Ю. Многозначность как прикладная проблема: семантическая разметка в национальном корпусе русского языка // Труды международной конференции «Диалог 2006». М., 2006. С. 445–450.
4. Ожегов С. И., Шведова Н. Ю. Толковый словарь русского языка : 80 000 слов и фразеологических выражений. М. : Азбуковник, 1999.
5. Поляков В. Н. Использование технологий, ориентированных на лексическое значение, в задачах поиска и классификации // Проблемы прикладной лингвистики. 2004. Вып. 2. С. 101–117.
6. Сиротина А. Ю., Лукашевич Н. В. Опыт создания корпуса текстов в сфере информационной безопасности // Сборник Международной научной конференции «Корпусная лингвистика-2019». СПб., 2019. С. 79–85.
7. Филаретова Л. П. Перспективы интегративной физиологии // Интегративная физиология. Всероссийская конференция с международным участием, посвящённая 95-летию Института физиологии им. И. П. Павлова РАН, Санкт-Петербург (9–11 декабря 2020 г.) : тезисы докладов. СПб. : Ин-т физиологии им. И. П. Павлова РАН, 2020. С. 3.
8. Шереметьева С. О., Бабина О. И. Платформа для концептуального аннотирования многоязычных текстов // Вестник ЮУрГУ. Серия «Лингвистика». 2020. Т. 17, № 4. С. 53–60.
9. Djema M., Candito M., Muller Ph., Vieu L. Corpus Annotation within the French Framenet: A Domain-by-Domain Methodology // Proceedings of the 10<sup>th</sup> International Conference on Language Resources and Evaluation. Portorož, Slovenia, 2016. P. 3794–3801.

10. Palmer M., Gildea P., Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles // *Computational Linguistics*. 2005. Vol. 31 (1). P. 71–106.
11. Rocha L., Soares-Bastos I., Freitas C., Rademaker A. Scavenger Hunt: What do we Find When Look for Confusions // *PROPOR: International Conference on the Computational Processing of Portuguese*. 2018. URL: [https://www.inf.ufrgs.br/propor-2018/wp-content/uploads/2018/10/PROPOR2018-SRW\\_paper\\_12.pdf](https://www.inf.ufrgs.br/propor-2018/wp-content/uploads/2018/10/PROPOR2018-SRW_paper_12.pdf) (дата обращения: 06.06.2022)
12. Sheremetyeva S. Automatic Extraction of Linguistic Resources in Multiple Languages // *Proceedings of NLPCS2012, 9<sup>th</sup> International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS2012, Wroclaw, Poland*. 2012. P. 44–52.
13. Viju J. S. Concept Interpretation by Semantic Knowledge Harvesting // *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*. 2018. Vol. 6 (5). P. 477–484.

## References

1. Efremova TF. *Novyj slovar' russkogo yazyka. Tolkovo-slovoobrazovatel'nyj = New Dictionary of the Russian Language. Explanatory and Word-Formation Dictionary*. Moscow, Russkij yazyk = Russian Language; 2001. 2354 p. (In Russ.).
2. Zinoveva A, Sheremetyeva S, Nerucheva E. The Analysis of Conceptual Ambiguity in Russian Texts. *Tyumen State University Herald. Humanities Research. Humanitates*. 2022;6(3(23)):38-60 (In Russ.).
3. Rakhilina EV, Kobritsov BP, Kustova GI, Lyashevskaya ON, Shemanayeva OJ. Semantic Ambiguity as an Application-Oriented Problem: Word Class Tagging in the RNC. *Proc. International Workshop Dialogue 2006*. 2006. Pp. 445–450 (In Russ.).
4. Ozhegov SI, Shvedova NYu. *Tolkovyj slovar' russkogo yazyka: 80 000 slov i frazeologicheskikh vyrazhenij = Explanatory Dictionary of the Russian Language: 80,000 words and phrases*. Moscow, Azbukovnik; 1999. (In Russ.).
5. Polyakov VN. Ispol'zovanie tekhnologij, orientirovannyh na leksicheskoe znachenie, v zadachah poiska i klassifikacii = Using Lexical Meaning-Oriented Technologies in Search and Classification Tasks. *Problemy prikladnoj lingvistiki = Issues of Applied Linguistics*. 2004;2:101-117 (In Russ.).
6. Sirotina AYu, Loukachevich NV. Opyt sozdaniya korpusa tekstov v sfere informacionnoj bezopasnosti = Experience in Creating a Text Corpus in the Field of Information Security. In: *Sbornik Mezhdunarodnoj nauchnoj konferencii "Korpusnaya lingvistika-2019" = Proceedings of the International Scientific Conference "Corpus Linguistics 2019"*. 2019. Pp. 79–85 (In Russ.).
7. Filaretova LP. Perspektivy integrativnoj fiziologii = The Prospects for Integrative Physiology. *Integrativnaya fiziologiya: Vserossijskaya konferenciya s mezhdunarodnym uchastiem, posvyashchyonnaya 95-letiyu Instituta fiziologii im. I. P. Pavlova RAN, Sankt-Peterburg (9–11 dekabrya 2020 g.) = Integrative Physiology: All-Russian Conference with International Participation Dedicated to 95<sup>th</sup> Anniversary of Pavlov Institute of Physiology Russian Academy of Sciences, St. Petersburg (December 9–11, 2020)*. 2020. P. 3 (In Russ.).
8. Sheremetyeva SO, Babina OI. A Platform for Knowledge Assisted Conceptual Annotation of Multilingual Texts. *Vestnik JuUrGU. Serija «Lingvistika» = Bulletin of the South Ural State University. Ser. Linguistics*. 2020;17(4):53-60. (In Russ.).
9. Djemaa M, Candito M, Muller Ph, Vieu L. Corpus Annotation within the French FrameNet: A Domain-by-Domain Methodology. In: *Proceedings of the 10<sup>th</sup> International Conference on Language Resources and Evaluation*. Portorož, Slovenia. 2016. P. 3794–3801.
10. Palmer M, Gildea P, Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*. 2005;31(1):71-106.
11. Sheremetyeva S. Automatic Extraction of Linguistic Resources in Multiple Languages. In: *Proceedings of NLPCS2012, 9<sup>th</sup> International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS2012*. 2012. Pp. 44–52.
12. Rocha L, Soares-Bastos I, Freitas C, Rademaker A. Scavenger Hunt: What do we Find When Look for Confusions. *PROPOR: International Conference on the Computational Processing of Portuguese*. 2018. URL: [https://www.inf.ufrgs.br/propor-2018/wp-content/uploads/2018/10/PROPOR2018-SRW\\_paper\\_12.pdf](https://www.inf.ufrgs.br/propor-2018/wp-content/uploads/2018/10/PROPOR2018-SRW_paper_12.pdf) (accessed 06.06.2022).
13. Viju JS. Concept Interpretation by Semantic Knowledge Harvesting. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*. 2018;6(5):477-484.

### Информация об авторе

**А. Ю. Зиновьева** — кандидат филологических наук, преподаватель кафедры лингвистики и перевода, лаборант научно-образовательного центра «Лингво-инновационные технологии».

### Information about the author

**Anastasia Zinoveva** — Candidate of Philological Sciences, lecturer, Department of Linguistics and Translation Studies, laboratory assistant, Research and Education Centre of Innovative Linguistic Technologies.

---

---

*Статья поступила в редакцию 26.05.2022; одобрена после рецензирования 11.06.2022; принята к публикации 26.12.2022.*

*The article was submitted 26.05.2022; approved after reviewing 11.06.2022; accepted for publication 26.12.2022.*

Автор заявляет об отсутствии конфликта интересов.

The author declares no conflicts of interests.