Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2023 Vol. 27 No. 2 468–499
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-32933
EDN: LOVEXS

Research article / Научная статья

Linguistic and statistical analysis of the lexical 'Langue-Parole' dichotomy in a restricted domain

Svetlana SHEREMETYEVA ✉ and Olga BABINA

South Ural State University, Chelyabinsk, Russia
✉sheremetevaso@susu.ru

Abstract
Development of new digital methods for analyzing the 'Langue-Parole' dichotomy is one of the most sought-after, but least researched problems of modern theoretical and applied linguistics. This determines the relevance of this study, the purpose of which is to develop a methodology for the automated linguastatistical analysis of a domain-related lexical layer in the context of the 'Langue-Parole' dichotomy and to apply the methodology to the Russian-language domain "Research on athlete integrative physiology" (RAIP). The study was conducted on the material of the Russian-language corpus including 56 RAIP domain texts of 300,000 wordforms in total published over the 2013–2020 period in the scientific journals "People. Sport. Medicine" (formerly "SUSU Bulletin. Series "Education, Healthcare, Physical Culture"), "Theory and Practice of Physical Culture", etc. The key methodological approach is the ontological analysis of corpus data using statistical and linguistic modeling methods. The domain-specific language and speech are modeled by the corresponding lexicon and corpus, while the 'Langue-Parole' lexical dichotomy is represented by the values of the linguistic-statistical concept verbalization parameters of the domain concepts in the lexicon and corpus. The computational parameters include the indices of lexical diversity, structural complexity, conceptual syncretism, lexical structural complexity vs. conceptual syncretism correlation, and syncretical concept junction when verbalized in the corpus. The main results of the study are: 1) a methodology for analyzing domain-specific lexical dichotomy 'Langue-Parole', which can be ported to other domains and national languages; 2) the RAIP domain-related resources, including language-independent ontology, conceptually annotated Russian corpus, onto-lexicon, linguistic-statistical parameter values of the lexical 'Langue-Parole' dichotomy; and 3) tools that automate certain stages of the study.
Keywords: *langue-parole dichotomy, linguastatistical analysis, restricted domain, ontology, Russian language*

For citation:
Sheremetyeva, Svetlana & Olga Babina. 2023. Linguistic and statistical analysis of the lexical 'Langue-Parole' dichotomy in a restricted domain. *Russian Journal of Linguistics* 27 (2). 468–499. https://doi.org/10.22363/2687-0088-32933

Лингвостатистический анализ лексической дихотомии «язык–речь» предметной области

С.О. ШЕРЕМЕТЬЕВА ✉, О.И. БАБИНА

Южно-Уральский государственный университет (национальный исследовательский университет), Челябинск, Россия
✉sheremetevaso@susu.ru

Аннотация
Создание новых информационно-технологических методов анализа соотношения языка и речи относится к числу наиболее востребованных, но наименее разработанных проблем современной теоретической и прикладной лингвистики, что определяет актуальность настоящего исследования, целью которого является разработка методологии автоматизированного лингвостатистического анализа лексического слоя ограниченной предметной области в контексте дихотомии «язык–речь» и реализация разработанных методологических принципов на материале русскоязычной предметной области «Исследование интегративной физиологии спортсменов» (ПО ИИФС). Исследование проводилось на материале русскоязычного корпуса из 56 текстов ПО ИИФС общим объемом 300 000 словоупотреблений, опубликованных в научных журналах «Человек. Спорт. Медицина» (ранее «Вестник ЮУрГУ. Серия «Образование, здравоохранение, физическая культура»), «Теория и практика физической культуры» и отдельных статей из сети Интернет за 2013–2020 гг. Основным методологическим принципом исследования является онтологический анализ корпусных данных с использованием лингвостатистических методов и методов лингвистического моделирования. Язык и речь предметной области моделируются соответствующими лексиконом и корпусом текстов, а специфика лексической дихотомии «язык–речь» определяется вычислением и сравнением значений лингвостатистических параметров вербализации концептуальной (онтологической) структуры предметной области в соответствующих лексиконе и корпусе. В качестве вычислительных параметров дихотомии предлагаются коэффициенты лексического разнообразия, лексической структурной сложности, лексического концептуального синкретизма, корреляции между структурной сложностью и концептуальным синкретизмом лексических единиц и сопряжения синкретично вербализованных концептов. Основными результатами исследования являются предложенная универсальная методология анализа лексической дихотомии «язык–речь», которая может быть использована в приложении к различным предметным областям и национальным языкам, и ориентированные на ПО ИИФС ресурсы: независимая от конкретного языка онтология предметной области, русскоязычные онто-лексикон, концептуально аннотированный корпус, значения лингвостатистических параметров лексической дихотомии «язык–речь» указанной предметной области и инструментарий, автоматизирующий определенные этапы исследования.
Ключевые слова: *дихотомия «язык–речь», лингвостатистический анализ, предметная область, онтология, русский язык*

Для цитирования:
Шереметьева С.О., Бабина О.И. Лингвостатистический анализ лексической дихотомии «язык–речь» предметной области. *Russian Journal of Linguistics.* 2023. Т. 27. № 2. С. 468–499. https://doi.org/10.22363/2687-0088-32933

1. Введение

Взаимодействие языка и речи как предмета теоретической и прикладной лингвистики, ориентированного на изучение структуры знания в соотношении с его вербализацией, является актуальной проблемой современности

(Осипова 2012) и требует новых информационно-технологических методов анализа дихотомиии «язык-речь», формализующих как процедуру, так и представление результатов анализа, чему до сих пор не уделяется достаточно внимания. Этим фактом определяется актуальность настоящего исследования, целью которого является 1) разработка методологии автоматизированого лингвостатистического анализа лексического слоя ограниченной предметной области в контексте дихотомии «язык–речь», 2) создание компьютерного инструментария для автоматизации этапов исследования и собственно вычислений параметров дихотомии, и 3) реализация разработанных методологических принципов на материале русскоязычной предметной области «Исследование интегративной физиологии спортсменов» (ПО ИИФС).

С одной стороны, представление о дихотомии «язык–речь» опирается на классические лингвистические идеи, сформулированные Ф. де Соссюром, который утверждал, что «речь является конкретной реализацией такого явления как язык, существующего в устной или письменной форме, причем «написанное слово, сливается с произнесенным до такой степени, что становится доминирующим»[1] (Saussure 1967: 27), и Л.В. Щербой, разделившим понятия «языковая система» – словари и грамматики, «речевая деятельность» – акты речи, и «языковой материал», который определяется как «совокупность всего говоримого и понимаемого в определенной конкретной обстановке» (Щерба 2004: 36). Отметим, что, по сути, уже Ф. де Соссюр подчеркивал роль текста как носителя речи, а в определении языкового материала Л.В. Щербы прослеживается то, что в современной лингвистике ассоциируется с понятием предметной области. При этом оба исследователя относят к языку некую константу, что, по формулировке Л.В. Щербы, фиксируется в словаре и грамматике.

Эти классические постулаты реализуются и в современных трактовках понятий «язык» и «речь», изменяясь со временем и отражая эволюцию развития многолетних теоретических и прикладных исследований и разработок, приобретая все больше оттенков интерпретации (Мельчук 1974). Язык понимается как определенная система знаков (код), включающая единицы разных уровней: фонетического, морфологического, лексического, синтаксического и семантического, а речь – как результат деятельности людей, использующих этот языковой код. Современная корпусная лингвистика открывает возможности для всестороннего изучения языка и речи как многоаспектных явлений, где понятие «речевой корпус», трактуется и как спектрограмма корпуса «звучащей речи», и как графическое (текстовое) представление речевых высказываний. При этом исследования, посвященные изучению функционирования вокабуляров словарей в корпусах текстов, по сути дела, рассматривают лексический уровень дихотомии языка и речи, и, как правило, предполагают использование статистических методов (Хохлова 2021, Шнякина 2015). При

[1] Оригинал цитаты: «...le mot écrit se mêle si intimement au mot parlé dont il est l'image, qu'il finit par usurper le rôle principal.» (Saussure 1967: 27).

этом в стремлении к получению конкретных результатов дихотомия «язык–речь» изучается на разных уровнях дискурса, например, на уровне подсистем лексики как в работе (Хохлова 2021), где сопоставляется словарно-корпусное функционирование атрибутивных коллокаций, или на лексико-концептуальном уровне, как, например, в исследовании, описанном в (Шнякина 2015), где автор рассматривает вербализацию концепта «событие». Подобные исследования проводятся как на материале общеупотребительного языка, так и на материале подъязыков ограниченных предметных областей (Чуфарова 2018). Последнее, очевидно, признает особый тип дихотомии, а именно, дихотомию "язык–речь" предметной области (ПО), которая имеет свою специфику. Выявление этой специфики имеет не только теоретическое, но и важное практическое значение, поскольку подавляющее число современных прикладных исследований в области реализации систем обработки информации, учитывающих реалистичность выполнения поставленных задач, создается с ориентацией на конкретные ПО. Важно отметить разнообразие современных методов прикладной лингвистики для анализа языкового материала (от собственно лингвистических, основанных на правилах, до статистических, нейросетевых и т.д.), которые, как правило, в качестве предварительного этапа требуют создания аннотированных корпусов текстов и инструментария, автоматизирующего обработку текстовой информации. Большое внимание уделяется разработке электронных словарей/лексиконов/баз данных и т.д. (Apresjan & Mikulin 2016). Электронные лексиконы часто создаются с удобными поисковыми модулями, а наиболее совершенные из них связаны с другими инструментами обработки естественного языка, что позволяет автоматизировать исследовательские процедуры (Tsalidis et al. 2004, Sheremetyeva 2018).

Настоящее исследование представляет собой попытку создать автоматизированный метод получения определенного представления о природе функционирования естественного языка в предметной области в контексте дихотомии «язык–речь» с помощью достаточно эффективного и объективного анализа. В частности, в работе описывается методология исследования лексической дихотомии «язык–речь» предметной области посредством автоматизированного лингвостатистического анализа лексико-концептуальной структуры словаря как модели языка и корпуса как модели речи. Конкретные этапы реализации предлагаемой методологии описываются на материале русскоязычной ПО «Исследование интегративной физиологии спортсмена» (ИИФС) и включают создание таких лингвистических ресурсов как онтология ПО, концептуально аннотированный корпус ПО и контентно-релевантный вокабуляр (словарь/лексикон) ПО. Для автоматизации создания ресурсов и собственно вычислений параметров дихотомии разработан специальный инструментарий, основными компонентами которого являются электронный лексикон, содержащий оцифрованные лексико-онтологические знания, и работающий на этих знаниях многоуровневый теггер.

Статья организована следующим образом. В Разделе 2 дается обзор современных подходов к аннотированию корпусов текстов на основе онтологий

как наиболее основополагающих для настоящего исследования. Раздел 3 представляет методологию исследования. Раздел 4 посвящен разработке ресурсов анализа, включающей в себя получение и формализацию знаний о предметной области, а также создание компьютерного инструментария. В Разделе 5 вычисляются лингвостатистические параметры вербализации концептов, характеризующие специфику лексической дихотомии «язык–речь» ПО ИИФС, и дается интерпретация значений вычисленных параметров. В Заключении приводится обзор основных результатов исследования их возможное применение.

2. Основные подходы к концептуальному аннотированию

2.1. Аннотирование и онтология

Проблемы аннотирования (разметки/тегирования) текстов находятся в центре внимания многих международных теоретических и прикладных лингвистических исследований, поскольку аннотирование является базовой процедурой любой последовательности шагов обработки текста, и его точность в значительной степени определяет качество конечных результатов исследований и разработок. Несмотря на то, что по умолчанию термин «аннотирование (тегирование)» чаще всего понимается как морфо-синтаксическая разметка лексических единиц, по мере того как эксплицитная семантика начинает играть все более заметную роль в компьютерных технологиях, ориентированных на интеллектуальную обработку неструктурированной информации, все чаще процедура проводится на концептуальном уровне (Stojanović et al. 2007). Понятие концептуальной аннотации обычно трактуется как определенный тип семантической аннотации, созданной для решения конкретных информационных задач в рамках конкретной предметной области, что, по сути дела, не совпадает с общепринятым пониманием семантической аннотации текста такими широкими семантическими признаками как «человек», «одушевленность» и т.п. Наборы концептуальных тегов (тегсетов), как правило, охватывают только контент, релевантный, прежде всего, для конкретной предметной области и задачи обработки информации. Тенденция к использованию концептуально аннотированных корпусов текстов в исследованиях, учитывающих семантико-концептуальные свойства естественного языка, особенно ярко проявляет себя с появлением общедоступного Интернета и популяризацией семантической паутины (Semantic Web).

В современных исследованиях, основанных на концептуальном аннотировании, выделяется два основных и пересекающихся направления: связывание сущностей (entity linking) и онтологический анализ. Связывание сущностей аннотирует лексические из текста соответствующими сущностями из целевой базе знаний, в качестве которой используется либо Википедия (Cucerzan 2007), либо, что становится более популярным, онтология (Galperin et al. 2022). В последнем случае процедура связывания является аналогом

онтологического анализа, который на практике заключается в аннотировании текстовых фрагментов тегами онтологических концептов и имеет достаточно серьезные недостатки (Gauch 2015). Трудно четко определить границы анализа, поскольку некоторые элементы текста могут не иметь онтологического отображения. Между теми текстовыми элементами, которые находят отражение в онтологии, и онтологическими концептами могут существовать отношения «многие-к-одному», «один-ко-многим» или «многие-ко-многим». Это ведет к концептуальной многозначности, разрешение которой бывает достаточно проблематичным. Универсального рецепта идеального онтологического анализа не существует, поэтому, как правило, в каждом практическом проекте разрабатываются специфические подходы к решению возникающих проблем. Далее, независимо от того, выполняется ли концептуальное аннотирование на основе онтологий вручную или автоматизированно (что является отдельной задачей), оно имеет еще одно очень серьезное ограничение – собственно наличие заранее построенной и устоявшейся онтологии, соответствующей целям исследования. Несмотря на то, что в настоящее время в открытом доступе находится довольно много онтологических библиотек, их пригодность для каждого конкретного исследовательского проекта, связанного с концептуальным аннотированием, как правило, проблематична. Поэтому в большинстве опубликованных работ, описывающих аннотирование на основе онтологического анализа, либо условно предполагается наличие необходимой онтологии (Ceausu & Després 2007), либо создание онтологического ресурса включается в проект в качестве предварительного этапа. При этом тип онтологии определяется языковым материалом, для аннотирования которого она создается. Обработка общих текстов требует наличия онтологий верхнего уровня, к которым относятся, например, онтологии Mikrokosmos (Nirenburg & Raskin 2004), SUMO (Niles & Pease 2003) и BFO (Arp et al. 2015), основной проблемой которых является покрываемость, поэтому в реальных проектах обработки и, в частности, концептуального аннотирования информации создаются более «узкие» онтологии предметных областей, достаточно хорошо покрывающие их концептуальную структуру и настроенные на выполнение конкретных задач обработки информации на одном (часто английском) или нескольких языках. К числу англоязычных относятся, например, онтология для анализа медицинских карт (Roberts et. al. 2009); онтология UMLS для аннотации медицинских терминов (Galperin et al. 2022), онтология предметной области «Терроризм» для предсказания терактов (Mannes & Golbeck 2005). Онтология для аннотирования русскоязычного корпуса ПО электронных услуг описана в работе (Добров и др. 2015).

В связи с большим количеством усилий и времени, необходимых для создания онтологий, в сферу исследовательского интереса попали многоязычные онтологии. При этом единого мнения относительно понимания многоязычия в онтологиях не существует. В рамках одного подхода онтологическое многоязычие трактуется как понятность (или адаптация) онтологических

меток для пользователей, говорящих на разных национальных языках. При другом подходе онтология считается многоязычной, если ее можно применять для обработки текстов на разных языках независимо от того, какой язык использовался для обозначения понятий. Эти интерпретации онтологического многоязычия напрямую зависят от определения онтологии как независимого или зависящего от конкретного языка ресурса.

Зависящие от конкретного языка онтологии, хорошо известным примером которых является знаменитая WordNet (Miller et al. 1990), представляют собой структуры, подобные тезаурусу, определяемые свойствами конкретного языка.

Переход к многоязычию трактуется как локализация онтологических понятийных обозначений. Подходы к самой локализации разнятся, а именно, локализация рассматривается как а) связывание значений слов разных национальных языков с онтологическими концептами посредством специально разработанной модели (Montiel-Ponsoda et al. 2008), б) перевод ярлыков (названий) онтологических концептов с одного языка на другой (Espinoza et al. 2008) и с) аннотацию онтологических концептов их названиями на разных языках вручную (Chaves & Trojahn 2010). Кроме того, ведутся исследования, связанные с универсальными методиками построения и/или обеспечения совместимости онтологий. Например, в (Alatrish et al. 2014) описано исследование, посвященное созданию универсальных инструментов для полуавтоматического построения одноязычных онтологий, а в (Embley et al. 2019) предлагается разработка открытых интерфейсов, обеспечивающих перекрестные ссылки на данные и метаданные одноязычных онтологий.

Что касается независимых от конкретного национального языка онтологий, таких как Mikrokosmos (Nirenburg & Raskin 2004), SUMO (Niles & Pease 2003) и BFO (Arp et al. 2015), они по определению допускают многоязычие в смысле способности обрабатывать тексты на разных языках, включая межъязыковое концептуальное аннотирование, что обеспечивается построением словарей конкретных языков и отображением значений их единиц в понятия одной и той же многоязычной онтологии. Этот трактовка многоязычности онтологии принята в настоящей работе.

2.2. Наборы тегов (тегсеты)

Проблема определения лучшего набора тегов для различных уровней аннотирования, включая концептуальное, имеет большое значение для корректности результатов, широко обсуждается в литературе и самым непосредственным образом связана с настоящим исследованием. При том, что большая часть дискуссий относительно тегсетов касается морфологического и синтаксического тегирования, их основные идеи релевантны и для концептуального аннотирования. (Elworthy 1995) при разработке набора тегов предлагает учитывать внешние и внутренние критерии. Внешний критерий требует,

чтобы теги могли кодировать различия в языковых характеристиках, которые требуются для задачи автоматической обработки текстов. Внутренний критерий дизайна тега касается максимальной точности процесса аннотирования. Считается, что меньший и более простой набор тегов должен повысить точность аннотирования, поскольку большой тегсет вызывает проблемы при создании надежных аннотационных инструментов. Однако дополнительная информация, включенная в тегсет, может помочь устранить возможную многозначность тега. (Nivre et al. 2008) утверждают, что точность аннотирования в решающей степени зависит от использования широкого спектра лингвистических характеристик, включая лексические. Таким образом, общепринятого решения проблемы «количество тегов» vs «точность тегов» не существует и в каждом конкретном исследовании принимается отдельное компромиссное решение.

При концептуальном тегировании дизайн тегсета во многом определяется размером и степенью детализации онтологии. (Carvalho et al. 2017) рассматривают онтологическую гранулярность с точки зрения онтологических уровней и предлагают сократить количество концептуальных тегов за счет использования определенных уровней так называемых многоуровневых онтологий. Другой широко используемый в настоящее время способ значительно сократить количество тегов и, тем не менее, воспользоваться преимуществами дополнительных лингвистических знаний описан в (Gnasa & Woch 2002), где для автоматизации процедур информационного поиска предложено использовать так называемые супертеги, кодирующие одновременно концепты предметной онтологии и синтаксические структуры. В общем случае, супертег может кодировать самый широкий спектр лингвистической информации в дополнение к концептуальной, что обеспечивает значительный выигрыш в производительности теггера.

Независимо от использования того или иного подхода тегсеты чаще всего создаются для определенного языка, и при изменении национального языка даже в пределах одной предметной области требуется новый набор тегов. Это исключает повторное использование ресурсов тегирования и негативно влияет на корректность, эффективность и совместимость исследований. Чтобы преодолеть эту проблему, были предприняты попытки разработать многоязычные универсальные тегсеты. Так, (Feldman et al. 2006, Erjavec 2010) сообщают о результатах экспериментов, проведенных на разных языковых семьях, и определяют наиболее сложные языковые явления, а (Petrov et al. 2012) предлагают использовать набор тегов из двенадцати крупных межъязыковых лексических категорий. Тем не менее, многие исследования указывают на то, что наборы тегов, разработанные для корпуса общей тематики, как правило, не годятся для корпусов предметных областей, и поэтому тегсеты следует ориентировать на конкретные предметные области и приложения (Orosz 2014).

3. Методология исследования и материал

Отметим, прежде всего, что важным аспектом нашей методологии исследования является ее ориентация на экономию затрат времени, исследовательских усилий и, как следствие, финансов, что предполагает максимально возможное повторное использование и адаптацию как ранее разработанных, так и созданных в процессе настоящего исследования ресурсов. При этом лингвистическая составляющая методологии основана на корпусных данных, что является основным приемом современных лингвистических исследований (Solovyev et al. 2022) и предполагает формализацию знаний предметной области на лексическом, морфо-синтаксическом и концептуальном уровнях с использованием количественных и качественных методов. Мы строго разделяем знания, зависящие от конкретного языка (в нашем случае — это лексикон) и знания, которые не зависят от какого-либо национального языка, а именно концептуальные. Для формализации последних используется онтология, которая в настоящем исследовании ориентирована на предметную область, и, следуя наиболее влиятельной трактовке, изложенной в работах (Nirenburg & Raskin 2004, Niles & Pease 2003, Arp et al. 2015), понимается следующим образом:

- Онтология – это независимый от языка ресурс, который служит посредником между одноязычными лексиконами.
- Онтология предметной области является неотъемлемой частью онтологии верхнего уровня, так как знания предметной области не изолированы от общих знаний о мире.
- Построение онтологий осуществляется в рамках индуктивно-дедуктивного подхода, начиная с предписанного набора концептуальных категорий, за которым следует их уточнение на основе корпусных данных.

Концептуальное знание, формализованное в виде независимой от конкретного языка онтологии предметной области, рассматривается в корреляции с его вербализацией в лексиконе и корпусе предметной области на конкретном языке, которые считаются моделями языка и речи ПО соответственно.

В качестве параметров, формализующих речевые связи лексической дихотомии «язык–речь» (более детальное описание и формулы для вычисления значений параметров приведены в разделе 5), предлагаются следующие:

- *коэффициент лексического разнообразия* – показатель, разработанный по аналогии с созданными ранее метриками (Варфоломеев 2000) и рассчитывается через соотношение между количеством лексических вербализаций концепта в тексте и количеством словоупотреблений в тексте.
- *коэффициент лексической структурной сложности* вербализации концепта показывает предпочтения по количеству компонентов в лексемах, значение которых описывается определенным концептом онтологии. Наша трактовка не предполагает учет морфологической структуры лексемы и/или типов словосочетаний.

- *коэффициент лексического концептуального синкретизма* – количественная оценка способности концепта быть вербализованным одной и той же лексической единицей одновременно с некоторыми другими, не противоречащими друг другу концептами, что отличается от качественных трактовок лексического синкретизма в (Пименова 2011, Сысоева 2019);
- *коэффициент корреляции между структурной сложностью и концептуальным синкретизмом* лексических единиц – количественная оценка зависимости между количеством компонентов лексической единицы и ее концептуальным значением, передаваемым одним или несколькими синкретичными концептами;
- *коэффициент сопряженности концептов* – количественный показатель, определяющий типичные комбинации концептов, синкретично вербализованных в лексических единицах корпуса.

Вычисление значений введенных параметров лексико-концептуальной дихотомии «язык–речь» предполагает использование концептуально аннотированного в процессе онтологического анализа корпуса ПО, специфика которого определяется следующим:

- Единицы аннотации включают в себя как однокомпонентные, так и многокомпонентные (до 10 слов) лексические группы, что более точно отражает контент ПО и снижает многозначность лексических единиц;
- Тегсет построен так, чтобы обеспечить баланс между теми признаками, которые имеют отношение к концептуальному аннотированию, и реалистичными ожиданиями их автоматического обнаружения;
- Концептуальное аннотирование предполагает выполнение двух последовательных процедур: автоматизированного онтологического анализа и устранения возможной концептуальной многозначности.

Ниже приведена основанная на изложенных выше методологических принципах дорожная карта исследования, предусматривающая несколько взаимосвязанных автоматизированных и выполняемых вручную этапов, к которым относятся:

1. Подготовка лингвистических данных:

- построение корпуса предметной области;
- анализ лексики и построение лексикона предметной области на морфосинтаксическом уровне;
- концептуальное структурирование лексики предметной области;
- построение онтологии ПО и онто-лексикона. Под онто-лексиконом понимается вокабуляр контентно-релевантных лексических единиц ПО, концептуальное значение которых описывается концептами онтологии. Построение новой онтологии предполагается в том случае, если не существует ресурса, покрывающего анализируемую ПО. В нашем случае наиболее близкими к ПО ИИФС являются биомедицинские онтологии, но даже наиболее разработанная из них, UMLS (Galperin et al. 2022) не содержит концептов, связанных со спортом, что обусловило создание нового онтологического ресурса,

- разработка инструментария для автоматизации процесса аннотирования;
- концептуальное аннотирование корпуса текстов посредством сочетания автоматических и полуавтоматических процедур.

2. Вычисление лингвостатистических параметров дихотомии «речь-текст» ПО.

Конкретная реализация изложенных методологических положений описывается в следующих разделах статьи на материале русскоязычной предметной области «Исследование интегративной физиологии спортсменов» (ПО ИИФС). Источником лингвистических данных является корпус текстов, включающий 56 русскоязычных научных статей предметной области «Исследование интегративной физиологии спортсменов», общим объемом 300 000 словоупотреблений. Корпус содержит опубликованные в интернете статьи из научных журналов за 2013–2020 гг. в области физиологии, медицины, физической культуры и спорта, таких как «Человек. Спорт. Медицина» (ранее «Вестник ЮУрГУ. Серия «Образование, здравоохранение, физическая культура»), «Теория и практика физической культуры» и прочих.

4. Подготовка исходных лингвистических данных

4.1. Анализ лексики

На первом этапе анализа лексики на основе корпуса ПО ИИФС составлен вокабуляр, единицами которого являются как одно-, так и многокомпонентные лексемы. Эта задача выполнена в два приема. Сначала корпус был обработан автоматическим экстрактором лексических групп длиной от 1 до 4-х компонентов (Sheremetyeva 2012), предварительно настроенного авторами статьи на русскоязычную ПО ИИФС. Затем в полученный список лексем были добавлены более длинные лексические группы (до десяти компонентов), полуавтоматически выявленные в корпусе с помощью функции «Найти». Из результирующего списка выделены две группы лексем: к первой группе отнесены общеупотребительные лексемы и лексемы, характерные для стиля научных статей. Во вторую группу лексем, включены единицы, отражающие информацию об объектах профессиональной деятельности исследователей ИИФС. Лексемы этой группы названы контентно-релевантными и проанализированы на морфо-синтаксическом уровне с вычислением статистического распределения типов лексики (см. рис. 1), результаты которого свидетельствуют о том, что основную нагрузку передачи контента ПО ИИФС несут именные группы. Последнее послужило основанием провести концептуальный анализ лексики ПО с целью выделения ее концептуальных классов, в первую очередь, на множестве контентно-релевантных именных групп, которые были разнесены в 39 концептуальных классов.

В качестве названий концептуальных классов использованы английские слова. Концептуальное значение каждого класса определяется не его названием, а исключительно дефиницией (см. табл. 1).

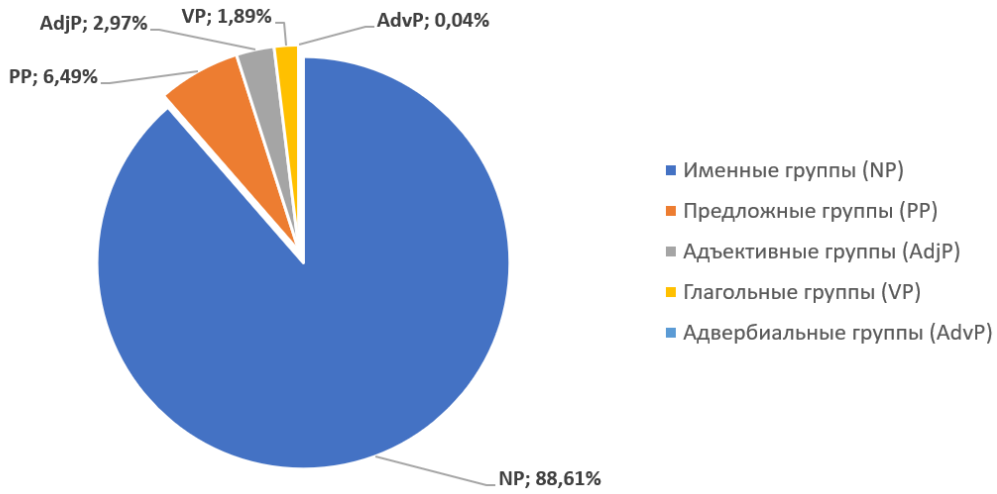

Рис. 1. Распределение типов лексических групп в корпусе ПО ИИФС /
Figure 1. Grammatical types of lexical units in the corpus of the RAIP domain

Таблица 1. **Фрагмент набора концептуальных классов контентно-релевантных лексем**

Концептуальный класс	Дефиниция	Примеры лексем класса
ATHLETE	Спортсмены и прочие лица, занимающиеся спортом (ЛЗС)	бегун-любитель, девушка-лыжница, мастер спорта, учащийся
ATHLETE-PHYSICS	Физические характеристики спортсменов и ЛЗС	выносливость, гибкость, подвижность суставов, устойчивость к стрессу
BODY-PART	Часть тела спортсменов и ЛЗС	верхние конечности, масса жировой ткани левой руки, нога
COMPETITION	Спортивное мероприятие	олимпиада, соревнование, чемпионат
ENVIRONMENT	Тип окружающей среды	атмосферное давление, равнина, среднегорье, уровень моря
EXAMINATION-METHOD	Методы обследования состояния спортсменов и ЛЗС	проба Ромберга с закрытыми глазами, спектральный анализ, стабилометрия
ORGANISM-BIOPROCESS	Биохимические процессы в организме спортсменов и ЛЗС	биосинтез, гликолиз, обмен веществ, экспрессия каспазы-32
TRAINING-PROCESS	Типы и процессы тренировочных нагрузок	общий объем нагрузки, тренировка, упражнение на гибкость

Table 1. **Set of domain-specific conceptual classes (fragment)**

Conceptual Class	Definition	Lexical Examples (Russian)
ATHLETE	Professional athletes and other people going in for sport	бегун-любитель, девушка-лыжница, мастер спорта, учащийся
ATHLETE-PHYSICS	Athletes' physical properties	выносливость, гибкость, подвижность суставов, устойчивость к стрессу
BODY-PART	Athletes' body parts	верхние конечности, масса жировой ткани левой руки, нога
COMPETITION	Sports events	олимпиада, соревнование, чемпионат
ENVIRONMENT	Type of environment where observation or training take place	атмосферное давление, равнина, среднегорье, уровень моря

Conceptual Class	Definition	Lexical Examples (Russian)
EXAMINATION-METHOD	Examination methods used in the research to measure athletes' states and parameters	проба Ромберга с закрытыми глазами, спектральный анализ, стабилометрия
ORGANISM-BIOPROCESS	Biochemical processes in athletes' organisms	биосинтез, гликолиз, обмен веществ, экспрессия каспазы-32
TRAINING-PROCESS	Training process-related concepts: types of workouts, training loads, etc.	общий объем нагрузки, тренировка, упражнение на гибкость

Набор концептуальных классов, выделенных на основании корпусного анализа контентно-релевантных именных групп, использован для классификации контенто-релевантных лексем остальных типов, что не потребовало введения новых классов.

4.2. Онтология и онто-лексикон ПО ИИФС

Концептуальные классы контентно-релевантной лексики, выделенные на предыдущем этапе анализа, выстроены в определенную систему отношений и приняты в качестве концептов и отношений онтологии ПО ИИФС, которая представлена в формализме онтологии MikroKosmos (Nirenburg & Raskin 2004). На рис. 2. приведен фрагмент разработанной онтологии, где три верхних уровня отражают деление мира, принятое в онтологии MikroKosmos, а выделенные жирным шрифтом концепты относятся к построенной нами двухуровневой предметной онтологии ИИФС.

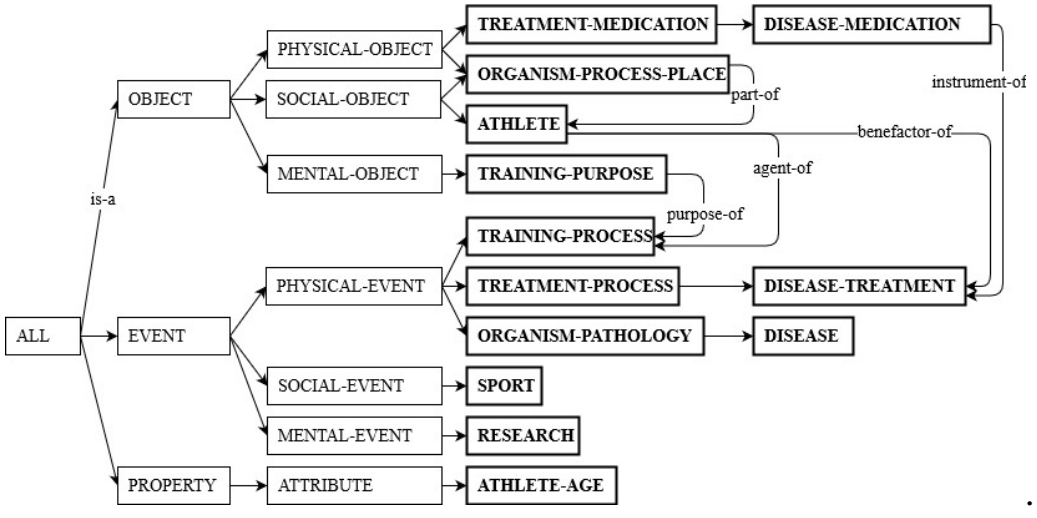

Рис. 2. Фрагмент онтологии ПО ИИФС /
Figure 2. Domain ontology for "Research on athlete integrative physiology" (fragment)

Отметим, что извлеченное указанным выше способом концептуальное знание, представленное в онтологии, не зависит от конкретного естественного языка и может быть повторно использовано для обработки текстов указанной предметной области на разных языках. Множество контентно-релевантных лексических единиц корпуса ПО, входящих в концептуальные классы, названо онто-лексиконом.

Важно отметить, что единица онто-лексикона может быть связана как с единственным концептом онтологии, т.е. быть концептуально однозначной, так и отображаться на несколько онтологических концептов и, следовательно, передавать несколько концептуальных значений. Последнее объясняется двумя различными лингвистическими явлениями: концептуальной многозначностью и концептуальной синкретичностью лексики.

Концептуально многозначными являются единицы онто-лексикона, имеющие различные противоречащие друг другу концептуальные значения. В каждом конкретном случае функционирования в корпусе ПО они реализуют только одно из этих значений. Например, лексема *кровь* концептуально многозначна, т.к. в корпусе может либо обозначать место, где происходят определенные физиологические процессы и таким образом реализовать концепт LOCALIZATION (*концентрация глюкозы в крови*), либо обозначать продукт метаболизма и реализовать концепт METABOLIC PRODUCT (*свойства крови*).

Концептуально синкретичными являются единицы онто-лексикона, одновременно реализующие несколько не противоречащих друг другу концептуальных значений. Примером концептуально синкретичной единицы онто-лексикона служит, например, трехкомпонентная лексема *время акклиматизации бегунов*, которая одновременно реализует концепты ATHLET, ORGANIZM REACTION, SPORTS and MEASURED PARAMETER. Чаще всего концептуальный синкретизм обнаруживается в многокомпонентных лексемах ПО. Однако, встречаются и однокомпонентные синкретичные лексемы. Например, единица онто-лексикона *лыжница* одновременно передает концептуальные значения ATHLETE, ATHLETE'S GENDER и SPORT.

На основе онтологии и результатов анализа лексики построен онто-лексикон ПО ИИФС с описанием морфо-синтаксических и концептуальных признаков лексических единиц, который принят в качестве модели лексического уровня языка ПО ИИФС. Единицей словаря считается словарная статья лексемы, связанной с одним концептом онтологии. Таким образом, концептуально синкретичные и многозначные лексемы онто-лексикона дают несколько словарных статей, количество которых соответствует числу концептов онтологии, с которыми связаны лексемы. Онто-лексикон оцифрован и занесен в программную оболочку разработанной платформы аннотирования.

4.3. Платформа аннотирования

В настоящем исследовании, как указывалось выше, кроме лексикона необходимым ресурсом вычисления параметров дихотомии «язык–речь» является концептуально аннотированный корпус ПО. В соответствии с используемой методологией, аннотирование основано на онтологическом анализе корпуса текстов, отображающем текстовые единицы на концепты онтологии предметной области. Это определило алгоритм реализации аннотационного процесса и архитектуру автоматизирующего этот процесс инструмента (платформы аннотирования). Программная оболочка платформы представляет собой значительно усовершенствованные и адаптированные для целей настоящего исследования модули, ранее созданные в рамках других проектов (Sheremetyeva 2018), что позволило значительно сократить затраты времени и усилий при выполнении основной цели настоящего исследования. Платформа аннотирования состоит из двух основных модулей: электронного лексикона и многоуровневого теггера. Программа платформы отделена от знаний и может быть использована для сбора и хранения лингвистических знаний на разных языках. Модули платформы снабжены интерфейсами с большим количеством функций, автоматизирующих процедуру сбора, поиска и проверки корректности знаний. Основным хранилищем лексических знаний является электронный лексикон, скриншот главной страницы интерфейса которого со знаниями ПО ИИФС дан на рис. 3.

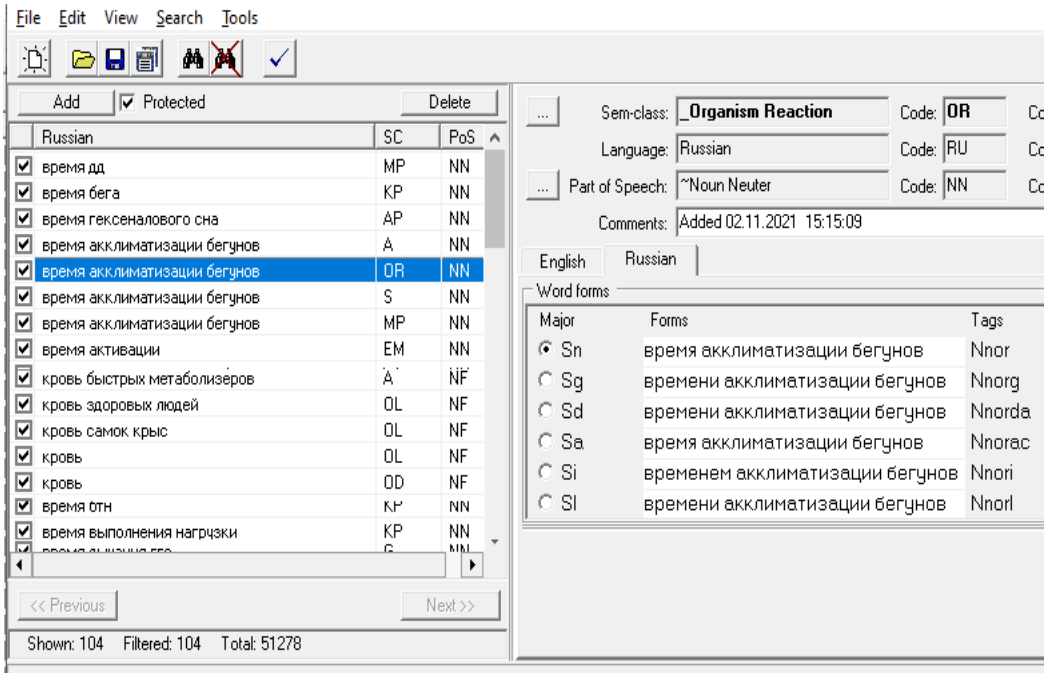

Рис. 3. Скриншот фрагмента главной страницы лексикона с открытой словарной статьей лексемы *время акклиматизации бегунов* /
Figure 3. Screenshot of the e-lexicon main page (fragment) with the lexicon entry for the Russian term '*время акклиматизации бегунов*'

Кнопки интерфейса не требуют пояснений. Все поля интерактивны и могут быть отредактированы. На левой панели даны (слева направо) интерактивный список единиц русского онто-лексикона ПО ИИСФ (щелчок по лексеме открывает ее словарную статью), соответствующие коды онтологических понятий (SC) и частей речи (PoS). Каждая словарная статья содержит лексему, связанную только с одним онтологическим концептом. Если лексема может быть отображена на несколько онтологических концептов, она появляется в разных словарных статьях электронного лексикона. Это объясняет лексические дубли в первой колонке. Например, на рис. 3 единица онто-лексикона *время акклиматизации бегунов* представлена в 4-х словарных статьях, поскольку она связана с 4-мя онтологическими концептами ATHLETE (A), ORGANIZM-RESPONCE (OR), SPORT (S) и MEASURED-PARAMETER (MP); выделена статья лексемы, связанная с онтологическим концептом ORGANISM-RESPONCE (OR). В центре правой панели интерфейса показана морфологическая зона словарной статьи с иконическим перечислением словоформ парадигмы лексемы. Соответствующие супертеги, кодирующие морфо-синтаксические и концептуальные признаки, показаны справа от словоформ парадигмы. Например, супертег Nnorg кодирует признаки N-существительное, n-средний род, g-родительный падеж, or-концептуальное значение ORGANIZM-RESPONCE.

Электронный лексикон подключен к теггеру, который настраивается на аннотирование с различной детализацией лингвистических характеристик: супертегами или только концептуальными тегами. В настоящем исследовании мы используем только последнее. Информация, кодируемая супертегами, является заделом на будущее и может быть использована, например, для разработки правил/метрик автоматического разрешения многозначных аннотаций, что не входит в задачи текущего проекта. Процедура концептуального аннотирования, включает два этапа:

1) поиск текстовых лексических единиц в морфологических зонах всех словарных статей лексикона и присвоение единицам онто-лексикона тегов всех связанных с ними онтологических концептов. Отметим, что явный (иконический) список словоформ парадигм лексем в словарных статьях лексикона позволяет избежать многих проблем морфологического анализа текста, например, необходимости автоматической лемматизации и связанных с ней ошибок;

2) разрешение концептуальной многозначности аннотаций и получение абсолютно корректно аннотированного («золотого») корпуса. Первый этап выполняется полностью автоматически, второй – в интерактивном режиме через интерфейс пост-редактора теггера. Экспериментальная проверка платформы аннотирования показала достаточность концептуального тегсета и значительное сокращение времени на получение «золотого» корпуса. Постановка, описание и анализ результатов эксперимента по тестированию платформы – тема отдельной статьи, которая готовится к печати.

На рис. 4 показан скриншот контрольного интерфейса теггера с автоматически аннотированным фрагментом корпуса ПО ИИФС. Значения концептуальных тегов: J = RESEARCH-AGENT, KR = TRAINING-PROCESS, MP = MEASUREMENT-PARAMETER, OP = ORGANISM-PHYSROCESS, OL = ORGANISM-PROCESS-PLACE, OR = ORGANISM-RESPONCE, ENVIRONMENT, W = TIME. Остальные теги аннотируют лексемы, не являющиеся контентно-релевантными.

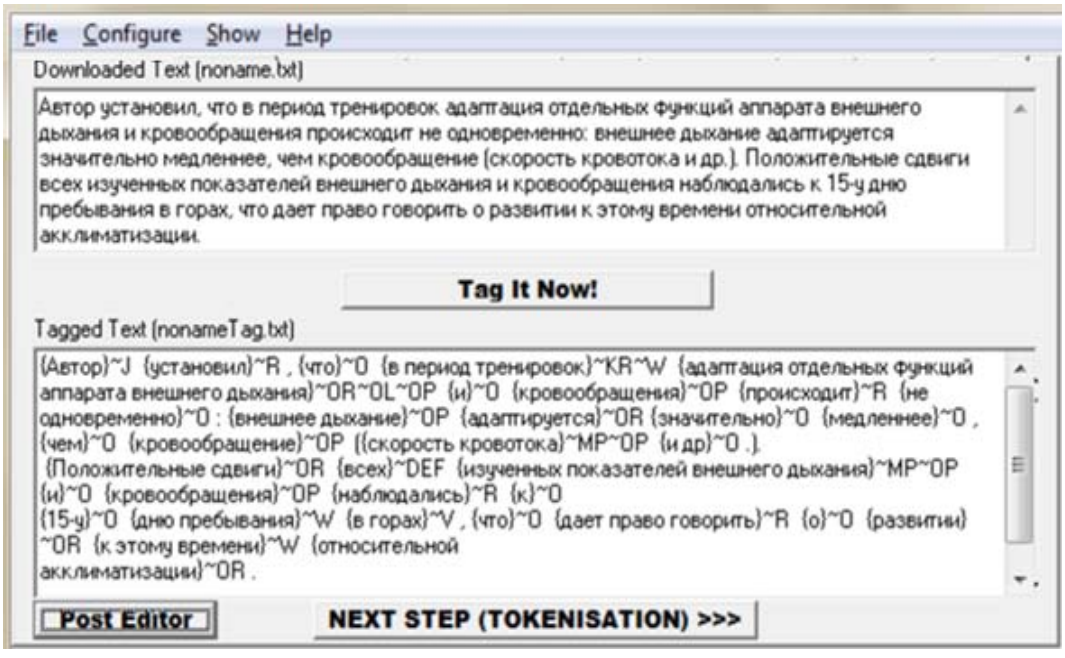

Рис. 4. Автоматически аннотированный фрагмент корпуса ПО ИИФС /
Figure 4. Automatically annotated corpus of the RAIP domain

5. Результаты и обсуждение

В настоящем разделе описаны процедура и результаты лингвостатистического анализа лексико-концептуальной дихотомии «язык–речь» ПО ИИФС на основе предварительно построенных и описанных выше лингвистических и программных ресурсов. Языковые параметры дихотомии вычислены на данных построенного онто-лексикона, моделирующего лексическую составляющую языка предметной области, который в настоящее время содержит 36269 словарных статей контентно-релевантных лексем. Для моделирования речи ПО ИИФС использован корпус**,** объемом 108030 словоформ, представляющий собой произвольно выделенную часть корпуса, использованного для моделирования языка (онто-лексикона) и онтологии. В нашем случае такой прием объясняется соображениями реалистичности выполнения задач исследования в обозримый период времени, поскольку, несмотря на автоматизацию определенных этапов работы, значительно ускоряющих исследование, качественное выполнение отдельных процедур может быть времязатратно.

Отметим, при этом, что в лингвистических исследованиях, основанных на сопоставлении «словарь-корпус», достаточно часто используются словари, вокабуляр которых по разным причинам построен не на анализируемом исследователями корпусе, как, например, в работе (Хохлова 2021).

Корпус, моделирующий речь ПО ИИФС, был сначала автоматически обработан с применением платформы аннотирования, что, в связи с наличием лексической многозначности и синкретичности, привело к появлению так называемых мульти-тегов, представляющих собой наборы тегов, поставленные в соответствие одной лексической единице. Затем многозначные мульти-теги были откорректированы в интерактивно-ручном режиме с помощью пост-редактора теггера. Полученный таким образом корректно аннотированный корпус назван «золотым» и далее использован для вычисления значений речевых параметров дихотомии.

Прежде всего путем автоматического анализа «золотого» корпуса инструментарием статистической обработки текстов вычислены реализованные в корпусе базовые для достижения цели исследования частотные характеристики единиц онто-лексикона и концептов онтологии, которые дают общую картину лексико-концептуальной структуры речи предметной области. В частности, определены частотные списки лексем, вербализующих в корпусе определенные концепты онтологии. В табл. 2 дан фрагмент наиболее частотных одно- и многокомпонентных единиц русскоязычного онто-лексикона ПО, концептуальное значение которых описывается онтологическими концептами ATHLETE, ORGANISM-PROCESS-PLACE и TRAINING-PROCESS.

Таблица 2. **Фрагмент наиболее частотных единиц онто-лексикона, вербализующих в корпусе концепты ATHLETE, ORGANISM-PROCESS-PLACE и TRAINING-PROCESS**

ATHLETE (f)	ORGANISM-PROCESS-PLACE (f)	TRAINING-PROCESS (f)
девушка (187)	мышца (122)	ДД (59)
юноша (152)	организм (73)	БТН (48)
спортсмен (136)	кровь (56)	основная стойка (41)
бегун (76)	МОК (49)	нагрузка (36)
учащийся (67)	миокард (40)	упражнение (34)
бегунья (53)	ткань (38)	тренировка (32)
подросток (51)	митохондрия (36)	прыжок в длину с места (24)
мужчина (36)	жировая ткань (31)	работа (24)

Table 2. **Top-frequent Russian onto-lexicon units verbalizing concepts ATHLETE, ORGANISM-PROCESS-PLACE and TRAINING-PROCESS**

ATHLETE (frequency)	ORGANISM-PROCESS-PLACE (frequency)	TRAINING-PROCESS (frequency)
девушка (187)	мышца (122)	ДД (59)
юноша (152)	организм (73)	БТН (48)
спортсмен (136)	кровь (56)	основная стойка (41)
бегун (76)	МОК (49)	нагрузка (36)
учащийся (67)	миокард (40)	упражнение (34)
бегунья (53)	ткань (38)	тренировка (32)
подросток (51)	митохондрия (36)	прыжок в длину с места (24)
мужчина (36)	жировая ткань (31)	работа (24)

Ниже приведены конкретные формулы и результаты вычисления значений описанных в разделе 3 параметров лексико-концептуальной дихотомии «язык–речь»**.**

Коэффициент лексического разнообразия вербализации концепта в корпусе текстов (уровень речи) и лексиконе (уровень языка) вычисляется по формулам (1) и (2) соответственно.

$$Var^{i}_{corpus} = \frac{n^{i}_{corpus}}{N}, \quad (1)$$

где Var^{i}_{corpus} – коэффициент лексического разнообразия *i*-го концепта в корпусе текстов; n^{i}_{corpus} – количество употреблений в корпусе одно- и многокомпонентных лексических единиц, вербализующих *i*-й концепт; *N* – количество употреблений единиц корпуса, вербализующих релевантные для данной ПО концепты;

$$Var^{i}_{lex} = \frac{n^{i}_{lex}}{V} \quad (2)$$

где Var^{i}_{lex} – коэффициент лексического разнообразия *i*-го концепта в лексиконе; n^{i}_{lex} – количество различных одно- и многокомпонентных лексических единиц, вербализующих *i*-й концепт, в лексиконе; *V* – количество лексических единиц в онто-лексиконе, вербализующих релевантные для данной ПО концепты.

Значения коэффициентов лексического разнообразия вербализации концептов в корпусе и лексиконе для наиболее частотных в корпусе концептов ПО ИИФС представлены в табл. 3. Список концептов ранжирован по убыванию значений коэффициента лексического разнообразия на корпусе; f^{i} – абсолютная частота в корпусе лексических единиц ПО, сопоставленных с *i*-м концептом.

Коэффициент корреляции Пирсона лексического разнообразия вербализации концептов в корпусе и лексиконе составляет 0,9733, что свидетельствует о высоком уровне положительной корреляции между этими показателями. Однако величина этих коэффициентов варьируется у различных концептов. Так, концепты MEASUREMENT, ATHLETE, TRAINING, EXAMINATION и RESEARCH демонстрируют высокое лексическое разнообразие как в корпусе, так и в лексиконе, что отражает основные направления исследований в области физиологии спортсменов.

Концепты со средним или низким коэффициентом лексического разнообразия в лексиконе и корпусе, тем не менее, могут быть достаточно частотными в корпусе, за счет высокой частоты вербализующих их онто-лексем, что и наблюдается при вербализации концептов SPORT и TIME. Это может свидетельствовать о том, что исследования в области физиологии спортсменов охватывают относительно небольшой набор видов спорта, при этом недостаточно исследований о взаимозависимости тренировок и временных периодов.

Таблица 3. **Коэффициенты лексического разнообразия наиболее частотных концептов**

Concept (*i*)	Var^i_{lex}	Var^i_{corpus}	f^i
ORGANISM	0,3470	0,2899	11552
MEASUREMENT	0,1478	0,1891	7535
ATHLETE	0,1261	0,1587	6323
TRAINING	0,0970	0,0807	3215
EXAMINATION	0,0718	0,0685	2731
RESEARCH	0,0576	0,0661	2635
SPORT	0,0233	0,0293	1168

Table 3. **Coefficients of lexical diversity for top-7 most-frequent concepts**

Concept (*i*)	Var^i_{lex}	Var^i_{corpus}	f^i
ORGANISM	0,3470	0,2899	11552
MEASUREMENT	0,1478	0,1891	7535
ATHLETE	0,1261	0,1587	6323
TRAINING	0,0970	0,0807	3215
EXAMINATION	0,0718	0,0685	2731
RESEARCH	0,0576	0,0661	2635
SPORT	0,0233	0,0293	1168

Коэффициент лексической структурной сложности вербализации концепта оценивается с помощью диаграммы размаха, построенной для каждого концепта и показывающей распределение по квартилям вербализующих концепт *n*-компонентных лексических единиц, $n \in [1 \dots 10]$, на основе данных о частоте таких единиц в корпусе и их количестве в онто-лексиконе.

Распределения лексико-структурной сложности вербализации концептов в лексиконе и корпусе графически представлены на рис. 5 и рис. 6 соответственно. На графиках утолщенная часть столбиков показывает значения n, на которые приходится второй и третий квартиль распределения количества *n*-компонентных лексических единиц, вербализующих концепт, в корпусе и лексиконе.

При этом для *n*, на которое выпала граница квартиля, учитывается доля *n*-компонентных единиц, попавших и не попавших в указанные квартили. Поэтому границы квартилей оказываются на графике смещены относительно целых значений *n*. Горизонтальная линия внутри утолщенного блока показывает медиану распределения. «Усы» на диаграмме показывают распределение первого и последнего децила лексико-структурной сложности вербализации концептов. Точкой на столбике обозначено средневзвешенное *n* для лексических единиц, вербализующих концепт.

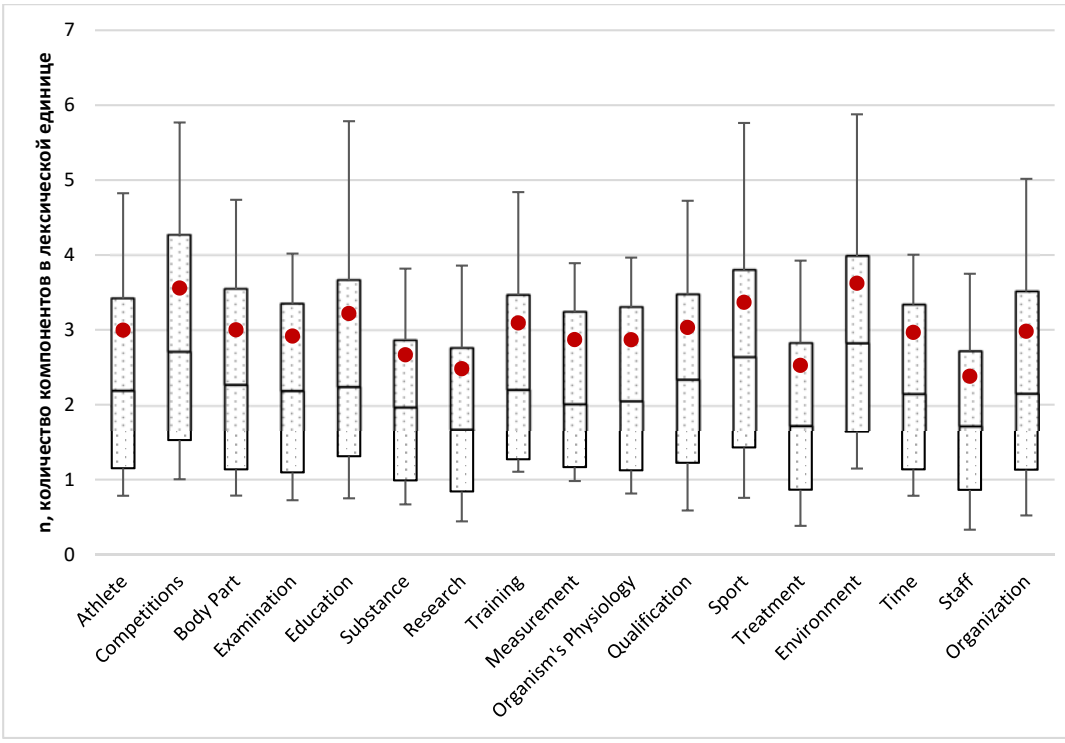

Рис. 5. Распределение структурной сложности вербализации концептов в лексиконе / Figure 5. Lexical structural complexity distribution per concept in the lexicon

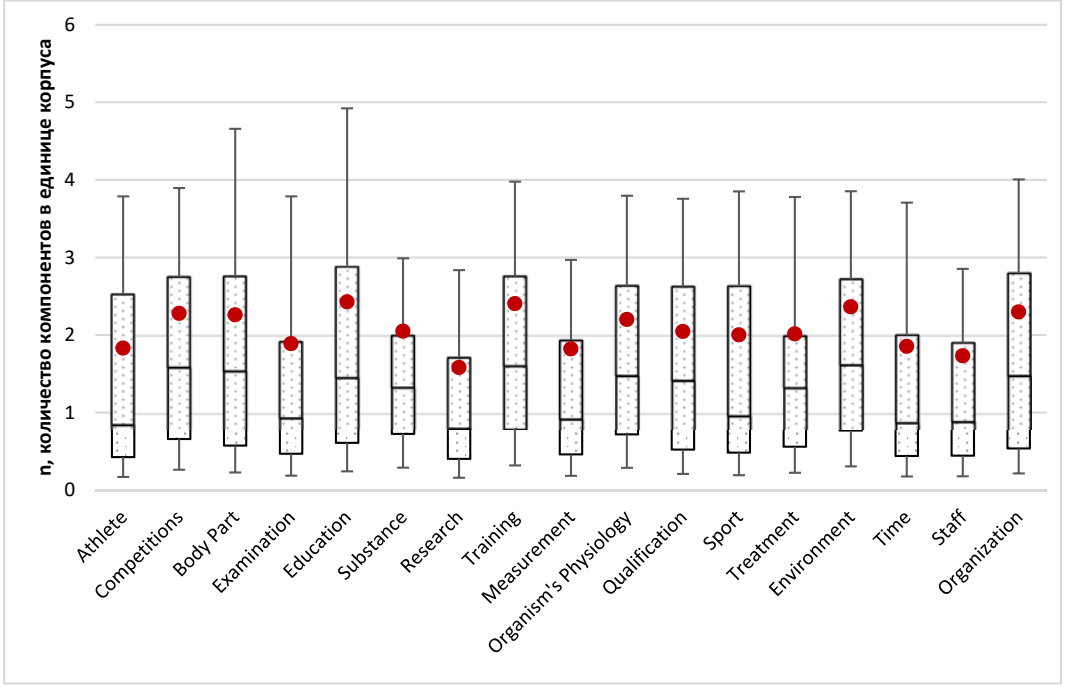

Рис. 6. Распределение структурной сложности вербализации концептов в корпусе / Figure 6. Lexical structural complexity distribution per concept in the corpus

Сопоставление распределений показывает, что вербализация концептов в корпусе имеют тенденцию к меньшей структурной сложности по сравнению с вербализацией аналогичных концептов в лексиконе. Длина многокомпонентных единиц в корпусе в 90% случаев не превышает 4 почти для всех концептов, лишь для концептов Body-Part и Education не превышает 5. В лексиконе лексические единицы имеют максимальную длину для различных концептов от 4 до 6. При этом медианные значения варьируют в корпусе от 0,8 до 1,6, в словаре находится в пределах от 1,6 до 2,8 для различных концептов. Для всех концептов и в корпусе, и в лексиконе медианные значения распределений ниже среднего *n*, что свидетельствует о правосторонней асимметрии в распределении лексических единиц по количеству компонентов (смещение в сторону меньшего количества элементов).

Распределение единиц с различной структурной сложностью в лексиконе достаточно равномерно: для каждого концепта величины второго и третьего квартиля не сильно отличаются друг от друга, варьируют около значения в 2 единицы, при этом медианные значения имеют тяготение к середине диапазона. Хотя абсолютные значения границ диапазонов отличается для различных концептов.

Распределение в корпусе характеризуется большей вариативностью. Вербализация концептов ATHLETE, EXAMINATION, RESEARCH, MEASUREMENT, SPORT, TIME, STAFF имеют тенденцию к структурной простоте (медианы этих концептов смещены вниз), эти концепты вербализуются в не менее чем половине случаев однокомпонентными единицами. В корпусе наблюдается бóльшая структурная однородность для одних концептов, так концепты EXAMINATION, SUBSTANCE, RESEARCH, MEASUREMENT, TREATMENT имеют наименьшую вариативность по параметру лексико-структурной сложности, представлены преимущественно одно-, иногда двухкомпонентными единицами. Другие концепты в большей степени вариативны, приблизительно в равной степени представлены одно-, двух- и трехкомпонентными единицами.

Коэффициенты концептуального синкретизма вычисляются по формулам (4) и (5) для единиц корпуса и лексикона соответственно.

$$Sync^{i}_{corpus} = \frac{s^{i}_{corpus}}{n^{i}_{corpus}} \quad (3)$$

где $Sync^{i}_{corpus}$ – коэффициент концептуального синкретизма вербализации *i*-го концепта в корпусе, s^{i}_{corpus} – частота единиц, соотнесенных одновременно с несколькими концептами, включая *i*-й концепт, в «золотом» корпусе; n^{i}_{corpus} – количество употреблений в корпусе однокомпонентных и многокомпонентных лексических единиц, соотнесенных с *i*-м концептом;

$$Sync_{lex}^{i} = \frac{s_{lex}^{i}}{n_{lex}^{i}} \quad (4)$$

где $Sync_{lex}^{i}$ – коэффициент концептуального синкретизма вербализации i-го концепта в лексиконе, s_{lex}^{i} – количество лексических единиц, допускающих множественное сопоставление с несколькими концептами ПО, включая i-й концепт; n_{lex}^{i} – количество различных одно- и многокомпонентных лексических единиц в лексиконе, сопоставленных с i-м концептом.

Значения коэффициентов концептуального синкретизма вербализации концептов ПО в корпусе и лексиконе представлены в табл. 4, f^{i} – абсолютная частота в корпусе лексических единиц, соотнесенных с i-м концептом.

Таблица 4. **Коэффициенты концептуального синкретизма единиц лексикона и корпуса**

Concept (i)	$Sync_{lex}^{i}$	$Sync_{corpus}^{i}$	f^{i}
ATHLETE	0,5257	0,6949	6323
SPORT	0,6995	0,6678	1168
ORGANISM	0,4403	0,4739	11552
MEASUREMENT	0,4681	0,3071	7535
TRAINING	0,2732	0,2802	3215
EXAMINATION	0,2219	0,1604	2731
RESEARCH	0,1370	0,0622	2635

Table 4. **Coefficients of conceptual syncretism for lexicon and corpus units**

Concept (i)	$Sync_{lex}^{i}$	$Sync_{corpus}^{i}$	f^{i}
ATHLETE	0,5257	0,6949	6323
SPORT	0,6995	0,6678	1168
ORGANISM	0,4403	0,4739	11552
MEASUREMENT	0,4681	0,3071	7535
TRAINING	0,2732	0,2802	3215
EXAMINATION	0,2219	0,1604	2731
RESEARCH	0,1370	0,0622	2635

Таблица 4 показывает, что около 2/3 единиц онто-лексикона, вербализующих концепт SPORT, и около половины случаев вербализации концептов ATHLETE, ORGANISM и MEASUREMENT, в действительности вербализуют более одного концепта. Эти результаты показывают концептуальную сложность терминологии ПО.

Коэффициент корреляции между структурной сложностью и концептуальным синкретизмом лексических единиц определяется в рамках проверки гипотезы о возможной функциональной зависимости между степенью структурной сложности лексической единицы и ее вербализацией с помощью нескольких синкретичных или только одного концепта. Для проверки гипотезы о возможной корреляции между вербализацией каждого концепта в лексиконе и корпусе мы построили таблицы сопряженности для переменных

«количество компонентов» в единице, вербализующей *i*-й концепт (*структурная сложность*), и «количество концептов, включая *i*-й, с которым(и) соотнесена эта единица (*концептуальный синкретизм*) в корпусе и лексиконе. Каждая переменная имеет два значения: «один компонент/концепт» и «несколько компонентов/концептов». Таким образом, таблицы сопряженности для вербализации концептов имели размерность 2x2. В ячейках таблицы сопряженности, основанной на лексиконе, указывалось количество различных единиц лексикона, вербализующих *i*-й концепт. Ячейки таблицы сопряженности на основе корпуса содержали корпусную частоту единиц, вербализующих *i*-й концепт. Примеры сопряженности по лексикону и корпусу для концепта TRAINING представлены в табл. 5.

Таблица 5. **Таблица сопряженности по корпусу и лексикону для концепта TRAINING**

	Лексикон		**Корпус**	
Концепты / Компоненты	**Один**	**Несколько**	**Один**	**Несколько**
Один	342	15	806	33
Несколько	463	111	649	137

Table 5. **Contingency table for the concept TRAINING**

	Lexicon		**Corpus**	
Concepts / Components	**One**	**Many**	**One**	**Many**
One	342	15	806	33
Many	463	111	649	137

Для определения степени корреляции между двумя переменными мы использовали редуцированный коэффициент корреляции, который на уровне лексики определяется как:

$$\phi^i_{lex} = \sqrt{\frac{{\chi^2}^i_{lex}}{n^i_{lex}}}, \quad (5)$$

где ϕ^i_{lex} – редуцированный коэффициент корреляции для измерения степени связи между структурной сложностью вербализации *i*-го концепта и множественной активацией концептов, включая *i*-й, единицей на уровне лексикона; ${\chi^2}^i_{lex}$ – значение статистики хи-квадрат для вербализации *i*-го концепта, найденное по таблице сопряженности для лексикона на основе списка извлеченных из корпуса лексических единиц, вербализующих концепты.

$$\phi^i_{corpus} = \sqrt{\frac{{\chi^2}^i_{corpus}}{n^i_{corpus}}}, \quad (6)$$

где ϕ^{i}_{corpus} – редуцированный коэффициент корреляции для измерения степени связи между структурной сложностью вербализации *i*-го концепта и множественной активацией концептов, включая *i*-й, единицей на уровне корпуса; ${\chi^2}^{i}_{corpus}$ – значение статистики хи-квадрат для вербализации *i*-го концепта, найденное для корпусной таблицы сопряженности на основе списка извлеченных из корпуса лексических единиц, вербализующих концепты.

Значения коэффициентов корреляции на основе корпуса и лексикона для 7 концептов, которые, по данным «золотого» корпуса, наиболее часто вербализуются синкретично с другими концептами, представлены в табл. 6.

Таблица 6. **Коэффициенты корреляции по сложности-синкретизму на основе корпуса и лексикона (фрагмент)**

Concept (i)	ϕ^{i}_{lex}	ϕ^{i}_{corpus}
ATHLETE	0,1974	0,0499
MEASUREMENT	0,2833	0,4855
TRAINING	0,2798	0,3109
ORGANISM	0,4051	0,4798
EXAMINATION	0,2939	0,3194
RESEARCH	0,2632	0,2582
SPORT	0,0396	0,0804

Table 6. **Correlation coefficients for association between lexical structural complexity and conceptual syncretism in the corpus and lexicon (fragment)**

Concept (i)	ϕ^{i}_{lex}	ϕ^{i}_{corpus}
ATHLETE	0,1974	0,0499
MEASUREMENT	0,2833	0,4855
TRAINING	0,2798	0,3109
ORGANISM	0,4051	0,4798
EXAMINATION	0,2939	0,3194
RESEARCH	0,2632	0,2582
SPORT	0,0396	0,0804

Согласно этим данным, мы наблюдаем незначительную положительную корреляцию между структурной сложностью единиц онто-лексикона и их способностью синкретично вербализовать несколько концептов. Отклонения от положительной корреляции между этими переменными могут быть вызваны следующими явлениями, наблюдаемыми в корпусе:

- однокомпонентные единицы, которые синкретично вербализуют несколько концептов. Так, концепт ATHLETE часто вербализуется онто-лексическими единицами, такими как *бегуньи, пловцы, лыжницы*, которые в дополнение к концепту ATHLETE синкретично вербализуют концепты ATHLETE-GENDER и SPORT;
- многокомпонентные онто-лексические единицы, которые состоят как из общих лексических единиц, так и из предметно-специфичных онто-лексических единиц. Например, в такой многокомпонентной онто-лексической

единице, как *специфические принципы спортивной тренировки*, вербализующей концепт TRAINING, к предметному онто-лексикону относятся только компоненты *спортивной тренировки*.

- многокомпонентные онто-лексические единицы, которые вербализуют определенное понятие ПО, при этом ни один из компонентов не принадлежит онто-лексикону (т.е. не является релевантным для ПО). Примером может служить онто-лексическая единица *сравнение уровней спектра в сагиттальной плоскости*, вербализующая концепт RESEARCH в теории и методике физической культуры и спорта.

Коэффициент сопряженности концептов, синкретично вербализованных в лексических единицах корпуса, мы строим на базе коэффициента сходства Жаккара (Jaccard 1901), рассчитываемым по формуле (7) с использованием данных по количеству употреблений вербализующих определенные концепты лексических единиц:

$$Jaccard_{corpus}^{i,j} = \frac{n_{corpus}^{i,j}}{n_{corpus}^{i} + n_{corpus}^{j} - n_{corpus}^{i,j}} \cdot \rho, \quad (6)$$

где $Jaccard_{corpus}^{i,j}$ – коэффициент степени связности концептов с синкретичной вербализацией (коэффициент Жаккара); $n_{corpus}^{i,j}$ – количество одно- и многокомпонентных единиц корпуса, соотнесенных одновременно с *i*-м и *j*-м концептами, n_{corpus}^{i} – количество единиц корпуса, соотнесенных с *i*-м концептом, n_{corpus}^{j} – количество единиц корпуса, соотнесенных с *j*-м концептом. Коэффициент ρ выбирался эмпирически кратным 10 для улучшения читаемости данных. В нашем случае мы взяли коэффициент 1 000.

Для наиболее частотных концептов, допускающих синкретичную вербализацию (AGENT, ORGANISM, SPORT, MEASUREMENT, TRAINING), вершина рейтинга наиболее тесно связанных синкретичных концептов в корпусе текстов приведена в табл. 7.

Таблица 7. **Рейтинг коэффициентов сопряженности синкретичных концептов в корпусе (фрагмент)**

Концепт (*i*)	Концепт (*j*)	$n_{corpus}^{i,j}$	$Jacc_{corpus}^{i,j}$	Примеры вербализации
AGENT	AGENT	1123	167,1877	*детей младшего звена, старшие лыжники*
AGENT	SPORT	627	144,7703	*представителей тяжелой категории борцов, лыжницы, стайер*
MEASUREMENT	ORGANISM	1289	93,3314	*повышенные значения митоза, показатели функции внешнего дыхания*
ORGANISM	ORGANISM	1084	68,7730	*метаболизм костной ткани, креатининкиназа сердца, повышенная продукция молочной кислоты*

Концепт (*i*)	Концепт (*j*)	$n_{corpus}^{i,j}$	$Jacc_{corpus}^{i,j}$	Примеры вербализации
AGENT	ORGANISM	378	31,5921	*снижение функционального состояния, сохранность высокой адаптивности*
ORGANISM	SUBSTANCE	274	30,6248	*ограниченный транспорт о2, истощение запасов калия*
MEASUREMENT	SUBSTANCE	219	30,1819	*индекс доставки кислорода, концентрация глюкозы*

Table 7. **Contingency coefficients for syncretic concepts in the corpus (fragment): in the last column examples of corpus units verbalizing syncretic concepts are given**

Concept (*i*)	Concept (*j*)	$n_{corpus}^{i,j}$	$Jacc_{corpus}^{i,j}$	Lexical Examples (Russian)
AGENT	AGENT	1123	167,1877	*детей младшего звена, старшие лыжники*
AGENT	SPORT	627	144,7703	*представителей тяжелой категории борцов, лыжницы, стайер*
MEASUREMENT	ORGANISM	1289	93,3314	*повышенные значения митоза, показатели функции внешнего дыхания*
ORGANISM	ORGANISM	1084	68,7730	*метаболизм костной ткани, креатининкиназа сердца, повышенная продукция молочной кислоты*
AGENT	ORGANISM	378	31,5921	*снижение функционального состояния, сохранность высокой адаптивности*
ORGANISM	SUBSTANCE	274	30,6248	*ограниченный транспорт о2, истощение запасов калия*
MEASUREMENT	SUBSTANCE	219	30,1819	*индекс доставки кислорода, концентрация глюкозы*

Представленный рейтинг степени связности синкретичных концептов наряду со значениями коэффициентов концептуального синкретизма могут служить основой для разработки метрик для автоматизации разрешения концептуальной многозначности при концептуальном аннотировании, например, путем исключения комбинаций преимущественно синкретичных тегов с высокой степенью связности из процедуры разрешения многозначности.

6. Заключение

Основными результатами проведенного исследования являются универсальная методология анализа лексической дихотомии «язык–речь», которая может быть использована в приложении к различным предметным областям и национальным языкам, методология разработки предметно-ориентированных онтологий; предложены методы разработки компьютерного инструментария аннотирования, а также разработаны конкретные ориентированные на ПО ИИФС ресурсы:

- программный инструментарий, автоматизирующий определенные этапы исследования;

- независимая от конкретного языка онтология ПО ИИФС,
- русскоязычный онто-лексикон ПО ИИФС,
- концептуально аннотированный «золотой» корпус ПО ИИФС,
- набор лингвостатистических параметров для количественной оценки связей между языком и речью предметной области на лексическом уровне,
- собственно вычисленные значения лингвостатистических параметров ПО ИИФС лексической дихотомии «язык–речь», демонстрирующие специфику лексической дихотомии русскоязычной ПО ИИФС.

Результаты исследования могут быть использованы для чтения курсов по компьютерной лингвистике, корпусной лингвистике, лексикографии и других прикладных дисциплин в области языкознания. Все вышесказанное позволяет утверждать, что представленная работа вносит определенный вклад в методологию изучения дихотомии «язык–речь» с использованием современных методов корпусной лингвистики, что делает их более объективными и дает возможность использования для создания компьютерных систем лингвистического анализа, Исследование вносит вклад в методику разработки онтологий, электронных лексиконов и компьютерных инструментов лингвистического анализа, специфика которых состоит в возможности применения разработанных методологических принципов к другим предметным областям и национальным языкам.

REFERENCES / СПИСОК ЛИТЕРАТУРЫ

Варфоломеев А.П. Психосемантика слова и лингвостатистика текста: метод. рекомендации к спецкурсу. Калининград: Калининградский университет, 2000. [Varfolomeev, Anatoly P. 2000. Psihosemantika slova i lingvostatistika teksta (Psychosemantics of the Word and Linguostatistics of the Text): Guidelines. Kaliningrad: Kaliningrad university Publ. (In Russ.)].

Добров А.В., Доброва А.В., Сомс Н.Л., Чугунов Н.Л. Семантический анализ новостных сообщений по теме «Электронные услуги»: опыт применения методов онтологической семантики // Государство и граждане в электронной среде: теория и технологии исследований. Труды XVIII объединенной конференции «Интернет и современное общество» IMS-2015. Санкт-Петербург: ИТМО, 2015. С. 120–125. [Dobrov, Aleksej V., Anastasija V. Dobrova, Nikolai L. Soms & Andrej V. Chugunov. 2015. Semanticheskii analiz novostnykh soobcshhenii po teme «Elektronnye uslugi»: opyt primeneniya metodov ontologicheskoi semantiki (Semantic analysis of news items on ‘electronic services’ subject domain: Experience of applying methods of ontological semantics). In *Gosudarstvo i grazhdane v ehlektronnoi srede: teoriya i tekhnologii issledovanii. Trudy XVIII ob’edinennoi konferentsii «Internet i sovremennoe obshchestvO» IMS-2015.* 120–125. Saint-Petersburg: ITMO Publ. (In Russ.)].

Мельчук И.А. Опыт теории лингвистических моделей Смысл ⇔ Текст: Семантика, синтаксис. 2-е изд. М.: Школа «Языки русской культуры», 1999. [Mel'chuk, Igor A. 1999. *On the Theory of Linguistic Models “Meaning ⇔ Text”.* 2nd ed. Moscow: Shkola «Yazyki russkoi kul'tury». (In Russ.)].

Осипова Л.И. К вопросу о дихотомии «язык–речь» // Актуальные проблемы гуманитарных и естественных наук. 2012. №11. С. 199–202. [Osipova, Lyudmila I. 2012.

K voprosu o dikhotomii “yazyk–rech’” (On the issue of the dichotomy “Langue-Parole”). *Aktual'nye Problemy Gumanitarnykh i Estestvennykh Nauk* 11. 199–202. (In Russ.)].

Пименова М.В. Лексико-семантический синкретизм как проявление формально-содержательной языковой асимметрии // Вопросы языкознания. 2011. № 3. С. 19–48. [Pimenova, Marina V. 2011. Leksiko-semanticheskii sinkretizm kak proyavlenie formal'no-soderzhatel'noi yazykovoi asimmetrii (Lexical and semantic syncretism as a manifestation of form- and content-related language asymmetry) // *Voprosy yazykoznaniya* 11. 19–48. (In Russ.)].

Сысоева А.А. Явление семантического синкретизма (на примере обозначений восприятия в немецком языке в диахронии) // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2019. Т. 817. № 1. С. 317–327. [Sysoeva, Alesia A. 2019. Yavlenie semanticheskogo sinkretizma (na primere oboznachenii vospriyatiya v nemeckom yazyke v diahronii) (Semantic syncretism (on the example of German lexical units denoting perception in diachrony)). *Vestnik Moskovskogo gosudarstvennogo lingvisticheskogo universiteta. Gumanitarnye nauki* 1 (817). 317–327. (In Russ.)].

Хохлова М.В. Атрибутивные коллокации в золотом стандарте сочетаемости русского языка и их представление в словарях и корпусах текстов // Вопросы лексикографии. 2021. № 21. С. 33–68. [Khokhlova, Maria V. 2021. Attributive collocations in the gold standard of Russian collocability and their representation in dictionaries and corpora. *Voprosy Leksikografii* 21. 33–68. (In Russ.)].

Чуфарова Е.Н. Юридический язык в дихотомии «язык–речь» // Юридические исследования. 2018. №2. С. 1–7. [Chufarova, Ekaterina N. 2018. Yuridicheskii yazyk v dikhotomii “yazyk–rech” (Legal language in the ‘language–speech’ dichotomy). *Yuridicheskie issledovaniya* 2. 1–7. (In Russ.)].

Шнякина Н.Ю. О вербализации событийных концептов // Историческая и социально-образовательная мысль. 2015. Т.7. № 5. Ч. 2. С. 283–288. [Shnjakina, Natal’ja Ju. 2015. O verbalizacii sobytiinyh konceptov (On event concept verbalization). *Istoricheskya i social'no-obrazovatel'naya mysl'* 7 (5–2). 283–288. (In Russ.)]. https://doi.org/10.17748/2075-9908-2015-7-5/2-283-288

Щерба Л.В. Языковая система и речевая деятельность. М.: Едиториал УРСС, 2004. [Scherba, Lev V. 2004. Yazykovaya sistema i rechevaya deyatel’nost’ (Language system and speech activity). Moscow: Editorial URSS. (In Russ.)].

Alatrish, Emhimed S., Dušan Tošić & Nikola Milenkov. 2014. Building ontologies for different natural languages. *Computer Science and Information Systems* 11 (2). 623–644. https://doi.org/10.2298/CSIS130429023A

Apresjan, Valentina & Nikolai Mikulin. 2016. Dictionary as an instrument of linguistic research, In Tinatin Margalitadze & George Meladze (eds.), *Proceedings of the XVII EURALEX international congress: Lexicography and linguistic diversity,* 224–231. Tbilisi: Ivane Javakhishvili Tbilisi State University.

Arp, Robert, Barry Smith & Andrew D. Spear. 2010. *Building Ontologies with Basic Formal Ontology*. Cambridge, MA: MIT Press.

Carvalho, Victorio A., Joo Paulo A. Almeida, Claudenir M. Fonseca & Giancarlo Guizzardi. 2017. Multi-level ontology-based conceptual modeling. *Data & Knowledge Engineering* 109 (C). 3–24.

Ceausu, Valentina & Sylvie Després. 2007. Learning term to concept mapping through verbs: A case study. *Proceedings of the Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM2007) located at the 4th International Conference on Knowledge Capture (KCap 2007), October 28-31*, *2007. CEUR Workshop Proceedings* 289. Whistler, British Columbia, Canada: CEUR-WS.org.

Chaves, Marcirio S. & Cassia Trojahn. 2010. *Towards a Multilingual Ontology for Ontology-Driven Content Mining in Social Web Sites.* https://www.researchgate.net/publication/266526035 (accessed 05 December 2022).

Cucerzan, Silviu. 2007. Large-scale named entity disambiguation based on Wikipedia data. *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, Prague, Czech Republic. 708–716. Association for Computational Linguistics.

Elworthy, David. 1995. Tagset design and inflected languages. In Steven P. Abney & Erhard W. Hinrichs (eds.), *Proceedings of the European chapter of the association for computational linguistics SIGDAT workshop from texts to tags: Issues in multilingual language analysis*, 1–10. Dublin: Association for Computational Linguistics.

Embley, David W., Stephen W. Liddle, Deryle W. Lonsdale & Yuri Tijerino. 2011. Multilingual ontologies for cross-language information extraction and semantic search. In Manfred A. Jeusfeld, Lois Delcambre & Tok Wang Ling (eds.), ER'11: *Proceedings of the 30th international conference on conceptual modeling*, 147–160. Berlin, Heidelberg: Springer-Verlag.

Erjavec, Tomaž. 2010. Multext-East version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the seventh conference on international language resources and evaluation* (LREC'10), 2544–2547. Valetta, Malta: European Language Resources Association.

Espinoza, Mauricio, Asunción Gómez-Pérez & Eduardo Mena. 2008. Enriching an Ontology with Multilingual Information. *The Semantic Web: Research and Applications. ESWC Lecture Notes in Computer Science* 5021. 333–347. Berlin, Heidelberg: Springer.

Feldman, Anna, Jirka Hana & Chris Brew. 2006. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of the fifth international conference on language resources and evaluation* (LREC 2006). 549–554. Genoa, Italy: European Language Resources Association.

Galperin, Rina, Shachar Schnapp & Michael Elhadad. 2022. Cross-Lingual UMLS Named Entity Linking using UMLS Dictionary Fine-Tuning. *Findings of the Association for Computational Linguistics: ACL 2022*. 3380–3390. Dublin, Ireland: Association for Computational Linguistics.

Gauch Jr, Hugh G. 2015. *Scientific Method in Practice*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511815034.

Gnasa, Melanie & Jens Woch. 2002. Architecture of a knowledge based interactive Information Retrieval System. *Proceedings of KONVENS 2002*. https://konvens.org/proceedings/2002/pdf/12P-gnasa.pdf (accessed 28 November 2022).

Hsieh, Hsiu-Fang & Sarah E. Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative Health Research* 15 (9). 1277–1288. https://doi.org/10.1177/1049732305276687

Jaccard, Paul. 1901. Étude comparative de la distribuition florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles* 37. 547–579. https://doi.org/10.5169/seals-266450.

Mannes, Aaron & Jennifer Golbeck. 2005. Building a Terrorism Ontology. In *Proceedings of the ISWC workshop on ontology patterns for the semantic Web* 36. https://www.semanticscholar.org/paper/Building-a-Terrorism-Ontology-Mannes-Golbeck/9bcb90e48677e39da7b84939e8c8da2b2a63cde7 (accessed 28 November 2022).

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* 3 (4). 235–244.

Montiel-Ponsoda, Elena, Guadelupe Aguado de Cea, Asunción Gómez-Pérez &Wim Peters. 2008. Modelling multilinguality in ontologies. In *Proceedings of COLING 2008, Companion volume – Posters*. 67–70. Manchester, UK: Coling 2008 Organizing Committee.

Niles, Ian & Adam Pease. 2003. Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology. In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*. 412–416.

Nirenburg, Sergei & Viktor Raskin. 2004. *Ontological Semantics*. Cambridge, MA: MIT Press.

Nivre, Joakim, Igor M. Boguslavsky & Leonid L. Iomdin. 2008. Parsing the SynTagRus treebank of Russian. In *Proceedings of the 22nd international conference on computational linguistics* (Coling 2008). 641–648. Manchester, UK: Coling 2008 Organizing Committee.

Orosz, György, Attila Novák & Gábor Prószéky. 2014. Lessons learned from tagging clinical Hungarian. *International Journal of Computational Linguistics and Applications* 5 (1). 129–145.

Petrov, Slav, Dipanjan Das & Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the conference on language resources and evaluation* (LREC 2012). 2089–2096. Istanbul, Turkey: European Language Resources Association.

Roberts, Angus, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts & Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics* 42 (5). 950–966.

Saussure, Ferdinand de. 1967. *Cours de Linguistique Générale*. Paris: Payot.

Sheremetyeva, Svetlana. 2012. Automatic extraction of linguistic resources in multiple languages. *Proceedings of NLPCS 2012, 9th International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS 2012, Wroclaw, Poland*, 44–52.

Sheremetyeva, Svetlana. 2018. Universal computational formalisms and developer environment for rule-based NLP. In Alexander Gelbukh (ed.), *Computational linguistics and intelligent text processing: CICLing 2017. Lecture notes in computer science* 10761, 67–78. https://doi.org/10.1007/978-3-319-77113-7_5

Solovyev, Vladimir, Marina M. Solnyshkina & Danielle M. McNamara. 2022. Computational linguistics and discourse complexology: Paradigms and research methods. *Russian Journal of Linguistics* 26 (2). 275–316. https://doi.org/10.22363/2687-0088-30161

Stojanović, Ljiljana, Nenad Stojanovic & Jun Ma. 2007. On the conceptual tagging: An ontology pruning use case. *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. 344–350.

Tsalidis, Christos, Aristides Vagelatos & Giorgos Orphanos A. 2004. An electronic dictionary as a basis for NLP tools: The Greek case, arXiv:cs/0408061 [cs.CL] (accessed 28 November 2022). https://doi.org/10.48550/arXiv.cs/0408061

Article history:
Received: 19 December 2022
Accepted: 15 May 2023

Сведения об авторах:
Светлана Олеговна ШЕРЕМЕТЬЕВА – доктор филологических наук, профессор, директор научно-образовательного центра «Лингво-инновационные технологии» ЮУрГУ. Имеет большой опыт преподавательской и научно-исследовательской работы в ЮУрГУ и за рубежом в качестве ведущего исследователя по вычислительной

лингвистике в университете Нью-Мексико (США) и лектора по вычислительной лингвистике и машинному переводу в Уппсалском университете (Швеция), а также Копенгагенской высшей школе экономики (Дания). Является регулярным участником, рецензентом и членом программных комитетов многих международных конференций по вычислительной лингвистике. Ее исследовательские интересы охватывают широкий круг проблем автоматической обработки текста.
e-mail: sheremetevaso@susu.ru
https://orcid.org/0000-0003-1245-4213

Ольга Ивановна БАБИНА – кандидат филологических наук, заведующий кафедрой лингвистики и перевода, заместитель директора научно-образовательного центра «Лингво-инновационные технологии» института лингвистики и международных коммуникаций Южно-Уральского государственного университета. Исследовательские интересы включают корпусную лингвистику, компьютерную лингвистику, автоматическую обработку текстов, применение методов машинного обучения для извлечения информации и интеллектуального анализа текстов.
e-mail: babinaoi@susu.ru
https://orcid.org/0000-0002-1733-6075

Bionotes:
Svetlana O. SHEREMETYEVA – Doctor Habil. in Computational Linguistics, Professor, Head of the Innovative Language Technology R&D center of the Institute of Linguistics and Intercultural Communication at the South Ural State University. She has considerable teaching and research experience acquired both in Russia and abroad. Prof. Sheremetyeva worked as a key researcher and lecturer in computational linguistics at New Mexico State University (USA), Uppsala University (Sweden), and Copenhagen Business School (Denmark). She is a regular participant, reviewer and program committee member of many international conferences on computational linguistics. Her research interests cover a wide range of NLP problems.
e-mail: sheremetevaso@susu.ru
https://orcid.org/0000-0003-1245-4213

Olga I. BABINA has a PhD in Computational Linguistics. She is Head of the Department of Linguistics and Translation and Deputy Head of the Innovative Language Technology R&D center of the Institute of Linguistics and Intercultural Communication at the South Ural State University. Her research interests include corpus linguistics, computational linguistics, natural language processing, as well as text mining and text analysis using machine learning methods.
e-mail: babinaoi@susu.ru
https://orcid.org/0000-0002-1733-6075