

Научная статья

УДК 81'322.2

DOI 10.25205/1818-7935-2023-21-2-89-104

Анализ расхождений в аргументационной разметке научных статей на русском языке

Иван Сергеевич Пименов

Новосибирский государственный университет
Новосибирск, Россия

Институт систем информатики им. А. П. Ершова СО РАН
Новосибирск, Россия

pimenov.1330@yandex.ru, <https://orcid.org/0000-0001-5946-9469>

Аннотация

В работе представлен анализ расхождений между аннотаторами при разметке аргументации в научных статьях, итогом которого является формулирование рекомендаций по стандартизации разметки. Расхождения между аннотаторами исследуются на трех уровнях разметки: при выявлении тезисов (утверждений в составе аргументов), построении связей между ними, определении типовых моделей рассуждения в основе связей. Исследуемым материалом служат 20 аргументационных аннотаций 10 научных текстов двух тематик, где для каждого текста построено два варианта разметки от разных экспертов. В совокупности 20 аннотаций содержат 917 тезисов и 773 аргумента. Разметка каждого текста заключалась в моделировании его аргументационной структуры согласно стандарту Argument Interchange Format, посредством построения ориентированного связного графа с двумя типами вершин: информационными, которые соответствуют тезисам, и вершинами-схемами, указывающими на связи между тезисами и реализуемыми в них типовые модели (схемы) рассуждения из компендиума Уолтона. Расхождения между аннотаторами выявляются посредством автоматического сопоставления графов, представляющих аргументационную структуру одного и того же текста, в три этапа: 1) определение тезисов, представленных в одном графе и отсутствующих в другом; 2) обнаружение связей, различным образом соединяющих одинаково выделенные тезисы; 3) выявление различных моделей рассуждения в одинаково построенных связях. Типовые случаи расхождений устанавливаются по итогам экспертного анализа найденных несоответствий между элементами графов с учетом их структурной специфики (расположения тезисов в графе, расстояния между связанными тезисами в исходном тексте, соотношения общих частот моделей рассуждения и числа вызванных ими расхождений). Показано, что соответствия между разметчиками достигают в среднем 78 % на уровне тезисов, 55 % для построения связей, 60 % для определения моделей рассуждения. К типовым случаям расхождения относится выявление тезисов, приводимых без обоснования; построение связей между тезисами в одном абзаце и на расстоянии более чем трех абзацев; выявление двух конкретных аргументационных схем (вызывающих 40 % и 33 % расхождений) и смешение функционально различных моделей рассуждения ввиду разноаспектного восприятия связей аннотаторами. Итогом исследования являются рекомендации по аргументационной разметке, сформулированные для минимизации типовых случаев расхождения на каждом уровне аннотирования.

Ключевые слова

разметка аргументации, расхождения в разметке, аргументационная структура текста, научные тексты, корпусная лингвистика, лингвистика текста

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 23-21-00325, <https://rscf.ru/project/23-21-00325/>.

Для цитирования

Пименов И. С. Анализ расхождений в аргументационной разметке научных статей на русском языке // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2023. Т. 21, № 2. С. 89–104. DOI 10.25205/1818-7935-2023-21-2-89-104

© Пименов И. С., 2023

Analyzing Disagreements in Argumentation Annotation of Scientific Texts in Russian Language

Ivan S. Pimenov

Novosibirsk State University
Novosibirsk, Russian Federation

A. P. Ershov Institute of Informatics Systems of SB RAS
Novosibirsk, Russian Federation

pimenov.1330@yandex.ru, <https://orcid.org/0000-0001-5946-9469>

Abstract

This paper presents the analysis of inter-annotator disagreements in modeling argumentation in scientific papers. The aim of the study is to specify annotation guidelines for the typical disagreement cases. The analysis focuses on inter-annotator disagreements at three annotation levels: theses identification, links construction between theses, specification of reasoning models for these links. The dataset contains 20 argumentation annotations for 10 scientific papers from two thematic areas, where two experts have independently annotated each text. These 20 annotations include 917 theses and 773 arguments. The annotation of each text has consisted in modelling its argumentation structure in accordance with Argument Interchange Format. The use of this model results in construction of an oriented graph with two node types (information nodes for statements, scheme nodes for links between them and reasoning models in these links) for an annotated text. Identification of reasoning models follows Walton's classification. To identify disagreements between annotators, we perform an automatic comparison of graphs that represent an argumentation structure of the same text. This comparison includes three stages: 1) identification of theses that are present in one graph and absent in another; 2) detection of links that connect the corresponding theses between graphs in a different manner; 3) identification of different reasoning models specified for the same links. Next, an expert analysis of the automatically identified discrepancies enables specification of the typical disagreement cases based on the structural properties of argumentation graphs (positioning of theses, configuration of links across statements at different distances in the text, the ratio between the overall frequency of a reasoning model in annotations and the frequency of disagreements over its identification). The study shows that the correspondence values between argumentation graphs reach on average 78% for theses, 55% for links, 60% for reasoning models. Typical disagreement cases include 1) detection of theses expressed in a text without explicit justification; 2) construction of links between theses in the same paragraph or at a distance of four and more paragraphs; 3) identification of two specific reasoning models (connected respectively to the 40% and 33% of disagreements); 4) confusion over functionally different schemes due to the perception of links by annotators in different aspects. The study results in formulating annotation guidelines for minimizing typical disagreement cases at each level of argumentation structures.

Keywords

argumentation annotation, inter-annotator disagreements, argumentation structure of a text, scientific texts, corpus linguistics, text linguistics

Acknowledgements

The research was conducted with the financial support of the Russian Science Foundation (project no. 23-21-00325, <https://rscf.ru/en/project/23-21-00325/>).

For citation

Pimenov I. S. Analyzing Disagreements in Argumentation Annotation of Scientific Texts in Russian Language. *Vestnik NSU. Series: Linguistics and Intercultural Communications*, 2023, vol. 21, no. 2, pp. 89–104. (in Russ.) DOI 10.25205/1818-7935-2023-21-2-89-104

Введение

Автоматический анализ аргументации (Argument Mining) является сравнительно молодой областью компьютерной лингвистики. Первая специализированная конференция по компьютерной обработке доказательств в текстах на естественном языке состоялась в 2014 году и обозначила начало активных исследований по развитию методов для таких задач, как выявление причин поддержки высказываемых мнений их авторами; обнаружение противоречий и разногласий в работах различных жанров; определение аргументационных ролей отдельных утверждений [Lawrence, Reed, 2019, С. 766–774]. Автоматический анализ организации рассуждений на уровне полных текстов проводится, например, для выявления тематической

или жанровой специфики отдельных работ [Al-Khatib, Wachsmuth, Hagen, Stein, 2017], оценки публикаций по убедительности выстраиваемых в них доказательств [Zagorulko, Domanov, Sery et al., 2020] или сложности этих доказательств для восприятия определенной аудиторией [Barbieri, Aggujaro, Molteni, Luzzatti, 2015], в том числе посредством кластеризации текстов по аргументационной сложности [Pimenov, Salomatina, 2022].

В свою очередь, разработка и экспериментальная проверка компьютерных методов аргументационного анализа требует использования корпусов текстов с разметкой структур рассуждения. Такая разметка, как правило, производится экспертами вручную либо с частичной автоматизацией и заключается в формальном представлении аргументационных структур отдельных текстов. Одной из моделей для формализации аргументов является стандарт AIF (Argument Interchange Format), согласно которому аргументационная разметка текстов производится поэтапно на трех уровнях: выделение тезисов, построение связей между этими тезисами, определение типовых моделей рассуждения при реализации каждой такой связи [Rahwan, Reed, 2009]. Выявление этих моделей предполагает обращение к классификации аргументационных схем, например, рекомендуемому авторами AIF компендиуму Дугласа Уолтона [Walton, Reed, Macagno, 2008], который содержит типовые модели рассуждения для дискурсов из разных областей (повседневного общения, научной коммуникации, юридических диспутов и других).

Однако создание аргументационных корпусов представляется особенно трудоемкой задачей ввиду требуемой многоаспектности разметки, а также лингвистической сложности аргументации как явления прагматического уровня языка. Как показано в [Котельников, 2018], сложность создания аргументационных корпусов замедлила исследования по Argument Mining на материале русскоязычных текстов. Одним из способов решения этой проблемы является разметка аргументации в отдельных утверждениях на уровне позиций, занимаемых их авторами по конкретным вопросам, и ролей приводимых доводов. Так, в [Kotelnikov, Loukachevitch, Nikishina, Panchenko, 2022] представлен аргументационный корпус из 9550 предложений (взятых из сообщений в социальных сетях), размеченных по отношению их авторов (поддержка, возражение, нейтральная позиция либо ее отсутствие) к трем близким темам, связанным с COVID-19: вакцинации, ношению масок, карантину.

Отмеченная сложность аргументационного аннотирования обуславливает важность формулирования строгих правил для уменьшения расхождений между разметчиками. В [Teguel, Cardellino, Cardellino et al., 2018] описывается анализ разногласий аннотаторов при разметке англоязычных юридических текстов по модели аргументов Стивена Тулмина, а результатом становится определение и исправление недостатков в изначальной методике аннотирования. Похожее исследование разногласий аннотаторов, но уже при разметке политических текстов на шведском языке по стандарту AIF и классификации схем Уолтона, представлено в [Lindahl, Borin, Rouses, 2019]: как показывают авторы, расхождения между разметчиками проявляются на всех трех уровнях аргументационной структуры (как при выделении тезисов, так и при построении связей, и при определении моделей рассуждения). Специфика аннотирования научных статей (англоязычных в области генетики) освещается в [Green, 2015]: разметчикам предлагалась классификация моделей рассуждения, разработанная с учетом характерных для генетических публикаций способов доказательства, однако исследование показало сложность выбора таких узкоспециализированных схем для аннотаторов без углубленной подготовки в области как генетики, так и лингвистики (среднее соответствие указанных ими схем с заданным образцом достигло 49 %). Наконец, в [Ilina, Kononenko, Sidorova, 2021] представлены принципы поэтапного аргументационного аннотирования на основе автоматического выявления маркеров, словарь которых составлен на корпусе научно-популярных статей на русском языке с учетом их жанровой специфики.

Таким образом, ввиду обозначенной сложности аргументационного аннотирования **целью представленного исследования** является выявление типовых расхождений между разметчи-

ками при моделировании аргументации в русскоязычных научных статьях. Для достижения этой цели предполагается решить следующие задачи:

- 1) выявить все случаи расхождения между аннотаторами на каждом уровне разметки посредством автоматической обработки аргументационных графов для отдельных научных статей и определения коэффициентов соответствия между вариантами разметки на этих уровнях;
- 2) определить типовые случаи рассогласования аннотаторов посредством частично автоматизированного и экспертного анализа расхождений в содержательном аспекте;
- 3) сформулировать рекомендации по уточнению принципов аргументационной разметки для предотвращения расхождений разметчиков в подобных типовых случаях.

1. Аннотирование аргументации согласно стандарту AIF

Моделирование аргументации на основе AIF подразумевает представление каждого текста в виде аргументационного графа с двумя видами вершин (информационными, обозначающими тезисы, и вершинами-схемами, указывающими на типовые модели рассуждения в основе каждого перехода между тезисами), где ребра между вершинами соответствуют аргументационным связям. На рис. 1 приведен пример фрагмента такого аргументационного графа, построенного посредством инструмента ArgNetBankStudio [Сидорова, Ахмадеева, Загоруйко и др., 2020].

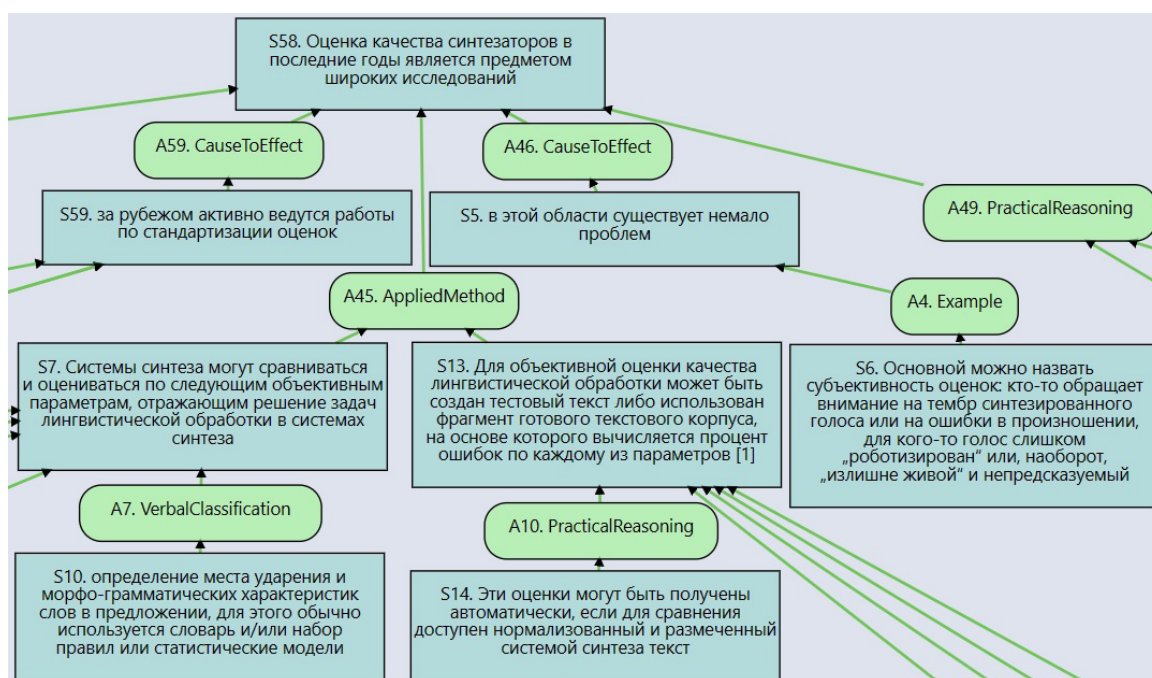


Рис. 1. Пример фрагмента аргументационного графа
Fig 1. An example of an argumentation graph fragment

Тезисы из исходного текста заключены в прямоугольные блоки и пронумерованы согласно порядку их представления в публикации. Стрелки у ребер указывают направления связей, а овальные блоки содержат типовые модели рассуждения (из компендиума Уолтона), реализующиеся при переходе от посылок к заключению. Роли утверждений определяются на уровне отдельных связей: один тезис может быть посылкой для другого, но при этом поддерживаться своими посылками, как S5 относительно S58 и S6. Вершина S58 соответствует главному тезису текста: к этой вершине существует путь по ребрам графа из любой другой (вся публикация

направлена на обоснование утверждения в S58), а из нее самой нет исходящих ребер (данный тезис не является посылкой ни для одного другого).

2. Количественная и качественная оценка согласия аннотаторов

Анализ единообразия аргументационной разметки заключается в сопоставлении разных вариантов построения аргументационных графов для одних и тех же текстов и производится в два этапа. На первом этапе автоматически подсчитывается количественная оценка соответствий между элементами аргументационной структуры отдельно на каждом из трех уровней разметки (выявление тезисов, построение связей между тезисами, определение типовых моделей рассуждения в основе этих связей). Второй этап подразумевает проведение экспертного анализа автоматически выявленных основных случаев расхождения между аннотаторами и определения их частот с учетом специфики элементов аргументации (структурной конфигурации тезисов в графе, протяженности связей между посылками и заключениями в линейном представлении текста, функциональной близости смешиваемых разметчиками схем). Такие элементы аргументационных структур, различающиеся между вариантами разметки, называются далее *спорными тезисами/связями/схемами*. Результаты второго этапа позволяют сформулировать рекомендации аннотаторам для повышения качества аргументационной разметки.

2.1. Поуровневая оценка соответствий между структурами аргументации

При выявлении и количественной оценке расхождений между аннотаторами коэффициенты соответствия подсчитываются для нескольких вариантов разметки отдельного текста посредством представленного ниже алгоритма.

Пусть $R = \{r_i^j\}$ – множество размеченных j экспертами ($j = 2$ в нашем случае, применение метода существенно не зависит от значения j) текстов ($i = 1, \dots, I$, I – число текстов в корпусе). Каждому варианту разметки r_i^j соответствует граф, который представляется множеством вершин $\{s_i^j\}$ и множеством соединяющих эти вершины помеченных ребер $\{e_i^j, ch(e_i^j)\}$, где e_i^j – ребра (связи), а соответствующие им метки $ch(e_i^j)$ – схемы аргументации.

Подсчет коэффициентов согласия (сходства графов аргументации) проводится последовательно в три шага: отдельно для $\{s_i^j\}$, подмножества $\{e_i^j\}$ и $\{ch(e_i^j)\}$. На первом шаге определяется доля утверждений $K(S_i^{12})$, совпавших в графах r_i^1 и r_i^2 , от суммарного числа аргументативных утверждений в этих графах: $K(S_i^{12}) = |S_i^{12}|/|\{s_i^1\} \cup \{s_i^2\}|$, где $S_i^{12} = \{s_i^1\} \cap \{s_i^2\}$. Под совпадением тезисов понимается точное вхождение одного в другой как подстроки: полное соответствие утверждений означает, что каждое из них является подстрокой другого, тогда как учет односторонних вхождений позволяет не рассматривать среди расхождений те пары тезисов, которые выделили оба разметчика при разном указании их точных границ (например, когда один аннотатор отсек вводный риторический маркер вида «во-первых, во-вторых...» либо убрал заключающий пунктуационный знак).

На втором шаге вычисляется доля совпавших связей от числа всех связей между вершинами из множества S_i^{12} : $K(\hat{E}_i^{12}) = |\hat{E}_i^{12}|/|E_i^{12}|$, где $\hat{E}_i^{12}$ – множество совпавших связей, а E_i^{12} – множество всех связей между вершинами из S_i^{12} .

На третьем шаге вычисляется доля совпавших меток (схем аргументации, или типовых моделей рассуждения) у выявленных на втором этапе связей из $\hat{E}_i^{12}$:

$$K(\hat{C}h_i^{12}) = |\hat{C}h_i^{12}|/|\hat{E}_i^{12}|, \text{ где } \hat{C}h_i^{12} \text{ – множество совпавших схем аргументации.}$$

Для наглядного представления коэффициентов согласия по всему корпусу подсчитываются диапазон их значений и среднее значение. Диапазон значений коэффициентов согласия вычисляется как $\min_i \{K(S_i^{12})\}$, $\min_i \{K(\hat{E}_i^{12})\}$, $\min_i \{K(\hat{C}h_i^{12})\}$ и $\max_i \{K(S_i^{12})\}$, $\max_i \{K(\hat{E}_i^{12})\}$, $\max_i \{K(\hat{C}h_i^{12})\}$ (соответственно минимальные и максимальные значения для каждого из трех

уровней разметки). В свою очередь, средние значения коэффициентов согласия для тезисов, связей и схем определяются следующим образом: $\sum_i K(S_i^{12})/I$, $\sum_i K(\hat{E}_i^{12})/I$, $\sum_i K(\hat{C}h_i^{12})/I$.

2.2. Анализ типовых случаев расхождения с учетом их специфики

По итогам количественной оценки коэффициентов соответствия между разметчиками выявлен общий набор случаев их расхождения на каждом уровне разметки. Уточнение анализа расхождений между аннотаторами предполагает выявление и подсчет частот отдельных типовых случаев рассогласования с учетом функциональной специфики соответствующих элементов аргументации. Дальнейшее описание метода ориентировано на случай с числом аннотаторов $j = 2$ ввиду вспомогательной роли коэффициентов соответствия для выявления расхождений в рамках данной работы.

В первую очередь, рассматриваются утверждения S^U , выделенные одним аннотатором и пропущенные другим (называются уникальными, U обозначает *Unique*): $S^U = \{S_i^U \mid i = 1, \dots, I\}$, где $S_i^U = (\{s_i^1\} \cup \{s_i^2\}) \setminus (\{s_i^1\} \cap \{s_i^2\})$. Для этих утверждений проверяется структурная специфика их реализации в аргументационном графе: определяются число утверждений S^L , обозначаемых листовыми вершинами в их содержащем варианте разметки (*Leaves*: эти вершины не имеют входящих связей, обозначаемые ими тезисы не поддерживаются явными посылками) и количество утверждений S^N (*Non-leaves*), которые обосновываются иными тезисами ($S^U = \{S^L\} \cup \{S^N\}$). Параллельно с разбиением S^U на S^L и S^N для каждого уникального утверждения $s^U \in S^U$ подсчитывается количество других уникальных тезисов среди непосредственно связанных с ним посылок и заключений, обозначаемое L^U (*Unique Linkages* – утверждения, смежные с данным в аргументационной структуре и пропущенные другим разметчиком). Определение частот типовых случаев расхождения на уровне тезисов предполагает подсчет количества утверждений, образующих S^L и S^N и их подгруппы, разграничиваемые по L^U .

При анализе расхождений при построении связей проверяются длины отдельных спорных связей $D_{p,c}$, где p обозначает посылку (*premise*), c – заключение (*conclusion*), а $D_{p,c}$ определяется на уровне абзацев как количество абзацных переходов между тезисами p и c в тексте (равняется 0, если утверждения находятся в одном абзаце). Анализ расстояний проводится на уровне абзацев, поскольку на этом уровне отражается авторское структурирование рассуждений (объединение нескольких предложений в целостный смысловой блок). Затем производится сравнение числа разногласий D_l для связей различных длин l .

Доля D_l для заданного l от суммарного числа расхождений для всех длин l обозначается в процентах как $R_l = D_l / F_l$, где F_l – общая частота связей длины l в корпусе. Анализ расхождений при связывании тезисов позволяет установить, насколько определенные длины связей влияют на разногласия разметчиков.

Автоматическая оценка различных типов расхождений может быть проведена и на уровне моделей рассуждения в основе аргументационных связей. Их анализ заключается в сравнении частот F_d между спорными парами схем: частота каждой такой пары соответствует числу случаев, когда один разметчик указывает одну из этих моделей, а другой аннотатор – другую (для всех $\hat{E}_i^{12}$). В свою очередь, подсчет показателей F_d дополняется следующими характеристиками: 1) долей текстов F_p , где $F_d > 0$, от числа всех текстов коллекции; 2) суммарной частотой F_D каждой отдельной схемы в аргументах с различным определением модели рассуждения между разметчиками (при одинаковом построении связей, определении заключения и посылок); 3) суммарной частотой F_a каждой отдельной схемы во всем корпусе (для проверки корреляции между частотой указания схемы разметчиками F_a и числа расхождений при ее выявлении F_D).

3. Анализ расхождений между разметчиками в аргументационном корпусе

Эксперимент по определению расхождений в аргументационной разметке проведен на корпусе из 10 научных статей, для каждой из которых подготовлено по два варианта разметки разными аннотаторами (по одному от каждого, ни один аннотатор не учитывал разметку второго). В совокупности корпус включает 917 аргументативных утверждений и 773 аргументов. Статьи корпуса принадлежат двум предметным областям (5 по компьютерным технологиям, 5 по лингвистике) и характеризуются объемом от 800 до 1400 слов (1077 слов в среднем). Оба разметчика специализируются в направлении компьютерной лингвистики (один является студентом 4 курса бакалавриата, другой – аспирантом 3 курса), в связи с чем обладают достаточной квалификацией для аннотирования текстов обеих выбранных областей. Данный корпус, на котором анализируются расхождения между разметчиками, является частью более крупного аргументационного корпуса из 34 научных статей для иных исследований (однако полный корпус содержит по одному выбранному варианту разметки для каждого текста).

Разметка научных статей корпуса проведена в соответствии со стандартом AIF с применением компендиума моделей рассуждения Уолтона. Всего использовано 19 различных моделей, передающих специфику организации рассуждений в научных статьях, 14 из которых соответствуют 98 % всех аргументов в корпусе. 10 наиболее частых схем отмечены в табл. 5. Процесс разметки регулирован набором дополнительных принципов, среди которых выделяются вспомогательные (начальная версия рекомендаций по определению тезисов и построению связей на основе авторской организации текста, по выявлению моделей рассуждения согласно бинарному дереву вопросов) и четыре ключевых:

- 1) для каждого текста определяется строго один главный тезис, выражающий основную идею научной статьи;
- 2) каждый текст представляется в виде связного, ациклического, ориентированного, корневого графа аргументации (корневой вершине соответствует главный тезис текста);
- 3) аргументативные утверждения и связи вводятся в граф итеративно от главного тезиса по расширяющейся окрестности (сперва определяются посылки к главному тезису, затем обоснования для этих посылок, и так до покрытия текста связной аргументационной структурой);
- 4) к аргументационному графу присоединяются только явно выраженные тезисы (без восстановления имплицитных утверждений, подразумеваемых автором, но не представленных в точной формулировке).

3.1. Количественная оценка расхождений на разных уровнях разметки

На первом этапе анализа расхождений между аннотаторами осуществлен подсчет коэффициентов их согласия по уровням разметки. Значения этих показателей в процентах приведены в табл. 1 (соответственно для тезисов, связей и аргументационных схем).

Таблица 1

Согласие аннотаторов на разных уровнях разметки

Table 1

Inter-annotator agreement on different annotation levels

Коэффициенты согласия	min значение	max значение	Среднее значение
$K(S_i^{12})$	62	87	78
$K(\hat{E}_i^{12})$	34	69	55
$K(\hat{Ch}_i^{12})$	41	82	60

Минимальные значения получены на тексте, в котором разметчики разошлись при выборе главного тезиса. Разногласия при определении главного тезиса сопряжены с расхождением в восприятии статьи, что отражается на всех уровнях разметки: потенциальные утверждения рассматриваются аннотаторами в разных аспектах, как направленные на доказательство различных идей, ввиду чего значительные противоречия проявляются и в построении связей, и в определении моделей рассуждения. Следует отметить, что во всех остальных текстах главные тезисы определены разметчиками одинаково.

Таким образом, оценка расхождений уже в поверхностном количественном аспекте указывает на важность согласования главного тезиса текста перед построением полных аргументационных графов для научных статей. Кроме того, как свидетельствуют средние значения коэффициентов согласия для разных уровней разметки, наиболее сложным этапом аннотирования выступает построение связей между тезисами (55 % совпадений в среднем по корпусе).

3.2. Характеристика основных типов расхождения между аннотаторами

При выявлении тезисов в научных статьях разногласия между аннотаторами проявились в 210 утверждениях из 917 во всем корпусе (23 %). Эти 210 утверждений возможно разделить по группам на основе двух критериев, обозначенных в 3.2 (наличия посылок и числа связанных тезисов, пропущенных другим аннотатором). В табл. 2 указано количество утверждений в каждой такой группе.

Таблица 2

Статистика расхождений по утверждениям разных групп

Figure 2

Frequencies of disagreements over statements of different types

Наличие посылок	L_U	Количество утверждений
S_L	1	49
	0	83
S_N	4-6	4
	2-3	14
	1	25
	0	34

Как показано в таблице, практически две трети спорных утверждений (83 + 49, 132 из 210) указаны без обоснований в содержащем их варианте разметки. Такие тезисы являются периферийными для аргументационных графов: если бы аннотатор, отметивший эти утверждения, тоже их пропустил, то его граф не претерпел бы значительных изменений (по сравнению с пропуском нелистовых утверждений, которые присоединяют свои посылки к главному тезису и при чем пропуске граф может утратить связность). Соответственно пропуск тезисов-листьев другим разметчиком не влечет расхождений в общей структуре аргументации (не отражается явно на иных тезисах, так как вершины-листья не присоединяют к графу других утверждений). Кроме того, 63 % из этих 132 утверждений (83) являются единственными пропусками в своей окрестности (все заключения, поддерживаемые ими, учтены и другим аннотатором). Стоит отметить, что не выявлено ни одного расхождения в тезисах-листьях, поддерживающих два и более заключения: такие посылки, приводимые без обоснования, но уточняющие несколько

других тезисов, нельзя рассматривать как периферийные, но именно при их выделении оба разметчика демонстрируют полное согласие.

Схожим образом почти для половины спорных тезисов, не являющихся листовыми, оба разметчика выявляют связанные с ними утверждения без дополнительных разногласий. Для трети утверждений из группы S_N аннотаторы расходятся лишь в одном присоединенном к ним тезисе, причем подобные утверждения характеризуются как минимум двумя связями (с заключением и с посылкой, поскольку не являются листовыми и направлены к главному тезису). С учетом обеих групп S_L и S_N лишь менее чем в 9 % случаев (18 из 210) разногласие аннотаторов по поводу одного тезиса сопряжено с их расхождением насчет двух и более тезисов, связанных с данным в содержащем его варианте разметки. Так, учет позиционной специфики спорных утверждений в аргументационных структурах позволяет уточнить их обобщенную количественную оценку $K(S_i^{12})$ и показывает, что разногласия между разметчиками проявляются в основном на уровне периферийных посылок и почти не затрагивают содержательно значимых тезисов. Ограниченная значимость этих посылок в аргументационной структуре и вызывает расхождение аннотаторов в их отделении от не относящихся к аргументации сегментов текста (не связанных с доказательством общей идеи).

На уровне связей отмечено 276 случаев разного соединения аннотаторами одинаково распознанных тезисов. В табл. 3 для каждой длины связи в абзацах l приведены абсолютная частота расхождений D_l по связям данной длины, относительная R_l от всех 276 расхождений и общая доля всех связей F_l в корпусе (как вызвавших разногласие, так и одинаково построенных).

Таблица 3

Расхождения при построении связей разных длин

Figure 3

Disagreements over links of different lengths

l	D_l	R_l	F_l
0	141	51 %	51 %
1	40	14 %	18 %
2	27	10 %	9 %
3	16	6 %	6 %
4-5	18	7 %	5 %
≥ 6	34	12 %	11 %

Как показывает таблица, почти 50 % всех расхождений проявляется при построении связей между тезисами в одном абзаце. На частоту этого типа спорных случаев влияет распространенность исходных связей в размеченных текстах, однако соединение тезисов внутри одного абзаца само по себе допускает вариативность, приводящую к разногласиям аннотаторов, несмотря на позиционную близость тезисов.

Иная ситуация наблюдается для соединения тезисов в соседних абзацах ($l = 1$). Хотя каждое седьмое разногласие разметчиков отмечается для утверждений в соседних абзацах, связи между ними покрывают почти пятую часть всего корпуса. Значения R_l и F_l оказываются близки и при $l = 2$, и при $l = 3$, что сближает разногласия на связях этой длины с расхождениями на связях внутри абзацев. Тем не менее, связи с $l = 0$ более проблематичны для разметки ввиду их высокой частоты в корпусе. Сложность построения связей возрастает для отдаленных те-

зисов ($l \geq 4$): такие связи соответствуют 16 % аргументам корпуса, но вызывают каждое пятое расхождение из-за значительного расстояния между тезисами.

Так, анализ расхождений на уровне связей в зависимости от их длины показывает, что основную сложность для разметки представляет соединение как предельно близких тезисов (внутри одного абзаца), так и наиболее отдаленных (на расстоянии четырех и более абзацев). Экспертный анализ спорных случаев позволяет выявить три ключевые причины, вызывающие большинство расхождений.

1) Детализация у одного разметчика промежуточного тезиса, отсутствующего у второго: аннотаторы могут согласиться в выявлении утверждений A и B , но при этом один из них построит связь вида $A \Rightarrow B$, а другой уточнит ее переходами $A \Rightarrow C$, $C \Rightarrow B$ (где C не выделен первым разметчиком). Соответственно, хотя оба разметчика распознали и тезис A , и тезис B , а также заметили связь между ними, один аннотатор уточнил этот логический переход через промежуточную посылку C (которую другой разметчик мог посчитать избыточной), и в его аргументационном графе прямая связь между A и B отсутствует.

2) Разное восприятие моделей рассуждения: разметчики по-разному определяют роли тезисов, а поэтому расходятся в их соотношении друг с другом.

3) Разная конфигурация связей в локальных блоках утверждений при указании одинаковых схем: тогда как соединения тезисов в разных абзацах более редки и, как правило, характеризуют ключевые для всего текста логические переходы между двумя явно логически связанными утверждениями, центральными для отдельных смысловых блоков, отдельный абзац может содержать несколько тезисов, соотношение которых в целостной структуре оказывается не таким ясным ввиду частного характера их содержания. В типичном случае таких расхождений оба разметчика выделяют тезисы A , B , C , где оба полагают C выводом из A и B , но один строит параллельные связи $A \Rightarrow C$ и $B \Rightarrow C$, а другой – последовательные $A \Rightarrow B$ и $B \Rightarrow C$.

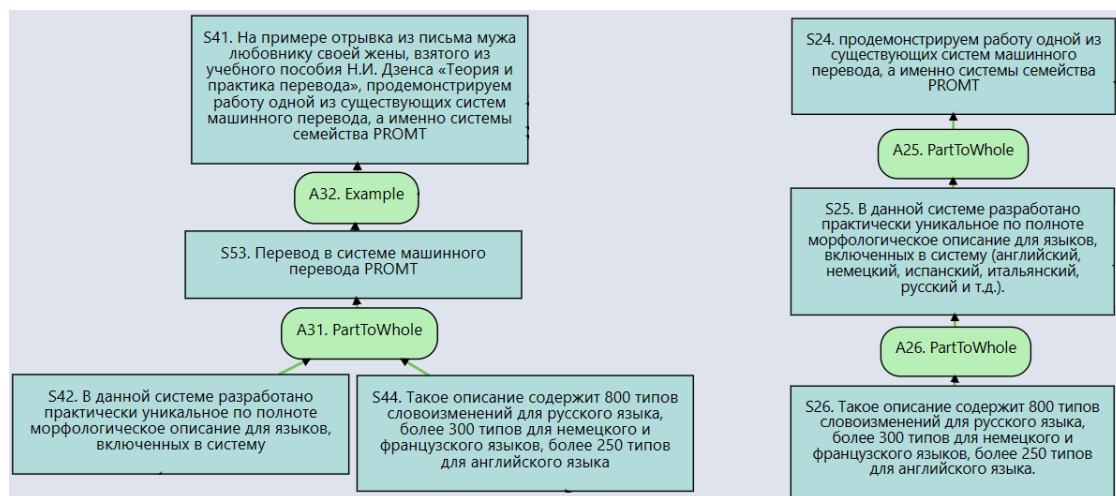


Рис. 2. Пример расхождения при построении связей
Fig. 2. An example of disagreement over links construction

На рис.2 приведен пример расхождения при построении связей внутри абзаца, в котором сочетаются все три указанные причины. Разметчики соглашались в выделении трех тезисов из одного абзаца (S41, S42 и S44 в варианте слева, и S24, S25, S26 в варианте справа), причем оба предполагают реализацию между ними модели Part To Whole. Однако один аннотатор добавляет промежуточный тезис S53 из другого абзаца (который непосредственно вводит пример перевода) и присоединяет к нему две посылки S42 и S44 параллельно, тогда как другой разметчик предполагает между ними последовательную связь (не считает их функционально

равноправными, а предполагает зависимость одной посылки от другой). Как следствие, хотя разметчики воспринимают фрагмент текста в достаточно близкой манере (оба выделяют три тезиса, выявляют среди них общий вывод на основе двух посылок), автоматическая проверка отмечает расхождение во всех пяти связях.

При выявлении моделей рассуждения отмечено 88 случаев, когда аннотаторы расходились во мнении о схеме при одинаковом распознавании тезисов и построении связей между ними. В табл. 4 приведены пары схем, вызвавшие разногласие разметчиков в наибольшем числе аргументов, с упорядочением по невозрастанию частоты встречаемости F_d ($F_d > 3$) и указанием доли текстов F_t , где эти расхождения проявились (от числа всех анализируемых текстов). Распределение схем пары между столбцами обусловлено алфавитным порядком.

Таблица 4

Частые пары схем, вызывающие расхождение разметчиков

Table 4

Frequent schemes pairs causing inter-annotator disagreement

№	Схема 1	Схема 2	F_d	F_t
1	Part To Whole	Verbal Classification	13	30 %
2	Part To Whole	Practical Reasoning	8	40 %
3	Cause To Effect	Part To Whole	6	30 %
4	Example	Verbal Classification	5	20 %
5	Cause to Effect	Practical Reasoning	4	20 %
6	Correlation to Cause	Sign	4	10 %
7	Practical Reasoning	Verbal Classification	4	10 %

Как показано в таблице, расхождения на 7 самых частых парах смешиваемых схем покрывают ровно половину всех разногласий на этом уровне разметки (44 из 88). Три самые частые пары смешиваемых схем включают модель Part To Whole. Важно отметить, что при анализе суммарных частот отдельных схем F_a (как в спорных случаях, так и при согласии разметчиков) не наблюдается прямого соответствия между данным показателем и частотой расхождений, вызываемых схемой. Иными словами, сложность определения модели обуславливается ее семантикой, а не частотой употребления. В табл. 5 приведены частоты F_a с уточнением их доли F_r относительно общего числа аргументов коллекции в процентах и суммарного числа расхождений F_D , связанных с этой моделью, для десяти схем с наибольшим значением F_a .

Как показано в таблице, самая частотная схема Part To Whole встречается в каждом пятом аргументе коллекции, однако связана почти с 40 % (34 из 88) всех расхождений на уровне схем. Другой моделью, отмечаемой в большом количестве расхождений (33 %, 29 из 88), является Verbal Classification, реализуемая в каждом седьмом аргументе. Именно две эти схемы чаще всего смешиваются разными аннотаторами. Экспериментальный подсчет $K(\hat{Ch}_i^{12})$ без учета модели Part To Whole показывает повышение среднего значения коэффициента с 60 до 67 %.

В свою очередь, модель Example употребляется в два раза чаще схем Correlation to Cause, Practical Reasoning и Sign, однако вызывает заметно меньшее число расхождений. Модель Sign, характеризующаяся довольно сложными принципами определения, близка схемам Correlation To Cause и Practical Reasoning по частоте, но вызывает почти в 1.5 раза меньше расхождений. Схожей частотой с этими тремя схемами характеризуется модель Expert Opinion, однако ее разметчики почти всегда указывают одинаково.

Таблица 5

Соотношение частот схем и вызываемых ими расхождений

Table 5

Frequencies of schemes and disagreements over them

№	Схема	F_a	F_r	F_D
1	Part To Whole	136	18 %	34
2	Verbal Classification	121	16 %	29
3	Example	95	12 %	13
4	Cause To Effect	89	11 %	19
5	Correlation To Cause	57	7 %	18
6	Practical Reasoning	51	7 %	18
7	Sign	45	6 %	11
8	Expert Opinion	43	6 %	4
9	Applied Method	27	3 %	3
10	Positive Consequences	21	3 %	5

3.3. Рекомендации по стандартизации аргументационной разметки

Таким образом, на основе количественного и качественного анализа расхождений между аннотаторами можно сформулировать следующие рекомендации для разметки аргументации.

1.1) На этапе выявления аргументативных утверждений представляется важным согласование главного тезиса между разметчиками. Расхождение в представлении главного тезиса влечет рассогласование всей аргументационной структуры текста (поскольку главный тезис обуславливает связность иных утверждений, направленных на его обоснование).

1.2) Согласование второстепенных тезисов (центров локальных блоков в аргументационном графе) представляется менее практичным. Как показывает анализ расхождений на уровне тезисов, они преимущественно связаны с выделением периферийных посылок (которые не оказывают особенного влияния на общую структуру разметки и смешиваются с неаргументативным содержанием текста).

1.3) Поскольку периферийные послылки, представленные в листовых вершинах графа, не так значимо влияют на общую аргументационную структуру, их согласование удобно производить уже после разметки (на дополнительном этапе постобработки). Такое согласование может производиться через сопоставление списков листовых вершин в двух вариантах аргументационного графа и унификацию множеств листовых вершин, для чего входящие в них вершины могут как отбрасываться, так и включаться, в зависимости от совместной экспертной оценки.

2.1) При построении связей между тезисами в одном абзаце желательно соблюдать его смысловую целостность: сперва определить главный тезис для этого абзаца, затем связывать утверждения из данного абзаца, не добавляя тезисы из других. Данная рекомендация подразумевает частичное ограничение одного из начальных принципов аннотирования: предлагается проводить разметку итеративно от главного тезиса уже с учетом абзацного членения текста (однако каждый отдельный абзац тоже размечать от его внутреннего главного тезиса к посыл-

кам). Когда все тезисы в абзаце связаны, присоединить его к аргументационному графу либо через исходящую связь от главного тезиса абзаца (не являющегося посылкой ни для одного утверждения в этом блоке), либо через входящую связь к одному из листьев абзаца (утверждению, не являющемуся заключением для других тезисов в этом смысловом блоке). Внешних связей, соединяющих данный абзац с остальным текстом, может быть несколько. Не следует нарушать смысловую целостность абзаца через введение внешних утверждений между его тезисами. Исключением являются случаи, когда ввиду авторской организации текста такое построение связей является единственным возможным вариантом разметки.

2.2) Если при построении связей внутри абзаца не получается определить его главный тезис, следует проверить, передает ли этот абзац частную идею, релевантную для развития аргументации в тексте. Если одно из утверждений в абзаце выражает эту идею, но его связи с другими тезисами абзаца являются неясными, то аргументативным следует указать только это утверждение, а другие утверждения абзаца убрать из списка аргументативных (не присоединять их к графу). Если такое утверждение не получается выявить, то следует рассматривать весь абзац как не относящийся к аргументации. Если абзац содержит несколько тезисов, релевантных для аргументации во всем тексте, но не связанных друг с другом, то следует присоединить каждое из этих утверждений к остальному графу изолированно от остальных (не рассматривать абзац как целостный смысловой блок), однако такое абзацное членение со стороны автора текста представляется нетипичным.

2.3) Если при построении связей внутри абзаца определен его главный тезис, но взаимные отношения других утверждений, поддерживающих этот тезис, представляются не совсем ясными, желательно присоединить эти утверждения к главному посредством параллельных связей. Такой стандартизирующий принцип позволяет повысить единообразие разметки и избежать частых расхождений при конфигурации связей.

2.4) Поскольку расхождения между разметчиками становятся сильнее с увеличением расстояния между тезисами в тексте, желательно избегать частого построения связей между позиционно отдаленными утверждениями (на расстоянии четырех и более абзацев). Каждую такую связь следует проверять на возможность ее разбиения на несколько связей меньшей длины через указание промежуточного тезиса. Однако если такой тезис в тексте отсутствует, либо не получается определить явную модель рассуждения для одной из двух новых связей (от исходной посылки к промежуточному тезису, от него к исходному заключению), то следует сохранить связь между отдаленными тезисами.

3.1) Для определения моделей рассуждения желательно уточнять не специфические особенности употребления отдельных схем (особенно конкретизированных, с относительно небольшой степенью абстрактности), а общие принципы выбора более крупных групп функционально схожих схем. Так, расхождения между разметчиками в этом исследовании связаны преимущественно с восприятием одних и тех же аргументов с разных сторон: смешиваются не близкие схемы, а функционально друг от друга далекие.

3.2) Поскольку основные расхождения между разметчиками на уровне схем связаны с разноаспектным восприятием связей (когда один или оба аннотатора замечают лишь одну возможность из нескольких и сразу указывают ее), желательно зафиксировать для разметки функциональную классификацию схем, предлагаемых аннотаторам. Каждую связь в отдельных аргументах следует тогда проверять на потенциальную реализацию схем из каждой функциональной группы. Обязательный учет всех функциональных групп для отдельных связей обеспечивает их многоаспектное восприятие аннотаторами, а группировка допустимых схем уменьшает количество вариантов для проверки.

Набор схем для разметки и их функциональную классификацию следует задавать с учетом жанровой и тематической специфики размечаемых текстов. В табл. 6 приведен пример классификации моделей рассуждения для научных статей, а сама классификация представлена в работе (Пименов, 2022).

Таблица 6

Пример функциональной классификации моделей рассуждения

Table 6

An example of functional classification of reasoning models

№	Функциональная группа	Модели рассуждения
1	Рассуждения практического характера	Practical Reasoning, Positive / Negative Consequences, Applied Method
2	Апелляции к авторитету	Expert Opinion, Popular Opinion, Position to Know
3	Причинно-следственные переходы	Cause to Effect, Correlation to Cause
4	Рассуждения через детализацию	Verbal Classification, Part to Whole, Example, Analogy, Sign

Заключение

По результатам анализа расхождений между аннотаторами при разметке научных статей на русском языке выявлены типовые случаи таких расхождений на трех уровнях разметки. Показано, что основные разногласия при выделении тезисов связаны с указанием частных посылок, приводимых в тексте без обоснования. Расхождения при выявлении иных тезисов, поддерживаемых иными и более значимых в аргументационной структуре текста, оказываются более редкими. Главные тезисы научных статей почти всегда определяются разметчиками одинаково, однако в единственном случае такого расхождения аннотаторы построили значительно различающиеся варианты разметки для одного текста, что подчеркивает важность предварительного согласования главного тезиса даже несмотря на отмеченную редкость расхождений.

На уровне связей между тезисами выявлены два основных типа расхождений: разметчики выстраивают разную организацию связей преимущественно между предельно близкими тезисами (представленными в тексте в одном абзаце) или, наоборот, значительно отдаленными (находящимися на расстоянии четырех и более абзацев). Первый случай вызван сложностью определения взаимных ролей между частными посылками, тогда как второй может быть частично решен указанием ограничений на допустимую длину связей. Наконец, при определении моделей рассуждения редки случаи смешения функционально близких схем, а основные расхождения проявляются в выборе более общих групп схожих моделей рассуждения.

В перспективе дальнейших исследований планируется проверка обозначенных рекомендаций по разметке аргументации при создании более крупного аргументационного корпуса из 100 научных статей.

Список литературы

- Котельников Е. В.** Извлечение аргументации из текстов и проблема отсутствия русскоязычных текстовых корпусов // *Advanced Science*, 2018, № 3 (11), С. 44–47.
- Пименов И. С.** Сочетаемость аргументов разных функциональных групп в научных текстах // *Филологические науки. Вопросы теории и практики*. Тамбов: Грамота, 2022. № 11. С. 3672–3680.
- Сидорова Е. А., Ахмадеева И. Р., Загорюлько Ю. А., Серый А. С., Шестаков В. К.** Платформа для исследования аргументации в научно-популярном дискурсе // *Онтология проектирования*, 2020, Т. 10 № 4 (38), С. 489–502.

- Al-Khatib K., Wachsmuth H., Hagen M., Stein B.** Patterns of Argumentation Strategies across Topics // *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September 7–11, 2017, pp. 1351–1357, 2017.
- Barbieri E., Aggujaro S., Molteni F., Luzzatti C.** Does argument structure complexity affect reading? A case study of an Italian agrammatic patient with deep dyslexia // *Applied Psycholinguistics*, Vol. 36, i. 3, 2015, pp. 533–558.
- Green N.** Identifying Argumentation Schemes in Genetics Research Articles // *Proceedings of the 2nd Workshop on Argumentation Mining*, Denver, 2015, pp. 12–21.
- Ilina D., Kononenko I., Sidorova E.** On Developing a Web Resource to Study Argumentation in Popular Science Discourse // *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”* (2021). Issue 20 (27). Moscow: RSUH, 2021. pp. 318–327.
- Kotelnikov E., Loukachevitch N., Nikishina I., Panchenko A.** RuArg-2022: Argument Mining Evaluation // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2022”*. 2022, pp. 1–16.
- Lawrence J., Reed C.** Argument Mining: A Survey // *Computational Linguistics*, Vol. 45, N 4, pp. 765–818, 2019.
- Lindahl A., Borin L., Rouces J.** Towards Assessing Argumentation Annotation – A First Step // *Proceedings of the 6th Workshop on Argument Mining*, Floreny, Italy, 2019, pp. 177–186.
- Pimenov I. S., Salomatina N. V.** Automatic Identification of Texts with Similar Argumentation Complexity // *2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*. Yekaterinburg, Russian Federation, 2022. pp. 810–815.
- Rahwan I., Reed C.** The argument interchange format // *Argumentation in artificial intelligence*, Rahwan I. and Simari G., Eds. Springer, 2009, pp. 383–402.
- Teruel M., Cardellino C., Cardellino F., Alemany L., Villata S.** Increasing Argument Annotation Reproducibility by Using Inter-annotator Agreement to Improve Guidelines // *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan, 2018.
- Zagorulko Yu. A., Domanov O. A., Sery A. S., Sidorova E. A., Borovikova O. I.** Analysis of the persuasiveness of argumentation in popular science texts // *Artificial Intelligence, Proceedings of the 18th Russian Conference RCAI*, pp. 351–367, September 2020.

References

- Al-Khatib K., Wachsmuth H., Hagen M., Stein B.** Patterns of Argumentation Strategies across Topics. In: *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September 7–11, 2017, pp. 1351–1357, 2017.
- Barbieri E., Aggujaro S., Molteni F., Luzzatti C.** Does argument structure complexity affect reading? A case study of an Italian agrammatic patient with deep dyslexia. *Applied Psycholinguistics*, Vol. 36, i. 3, 2015, pp. 533–558.
- Green, N.** Identifying Argumentation Schemes in Genetics Research Articles. In: *Proceedings of the 2nd Workshop on Argumentation Mining*, Denver, 2015, pp. 12–21.
- Ilina, D., Kononenko, I., Sidorova, E.** On Developing a Web Resource to Study Argumentation in Popular Science Discourse. In: *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”* (2021). Issue 20 (27). Moscow: RSUH, 2021. Pp. 318–327.
- Kotelnikov, E. V.** Extraction of argumentation from texts and the problem of lack of corpora with texts in Russian language. *Advanced Science*, 2018, vol. 3 (11), pp. 44–47. (in Russ.)
- Kotelnikov, E., Loukachevitch, N., Nikishina, I., Panchenko, A.** RuArg-2022: Argument Mining Evaluation. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2022”*. 2022, pp. 1–16.

- Lawrence, J., Reed, C.** Argument Mining: A Survey. *Computational Linguistics*, 2019, vol. 45, n 4, pp. 765–818.
- Lindahl, A., Borin, L., Rouces, J.** Towards Assessing Argumentation Annotation – A First Step. In: *Proceedings of the 6th Workshop on Argument Mining*, Florency, Italy, 2019, pp. 177–186.
- Pimenov, I. S.** Compatibility of Arguments from Different Functional Groups in Scientific Texts. *Philology. Theory & Practice. Tambov: Gramota*, 2022, vol. 11. pp. 3672–3680. (in Russ.)
- Pimenov, I. S., Salomatina, N. V.** Automatic Identification of Texts with Similar Argumentation Complexity. In: *2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*. Yekaterinburg, Russian Federation, 2022. pp. 810–815.
- Sidorova, E. A., Akhmadeeva, I. R., Zagorulko, Yu. A., Sey, A. S., Shestakov, V. K.** Research platform for the study of argumentation in popular science discourse. *Ontology of Designing*, 2020, vol. 10 (4). pp. 489–502. (in Russ.)
- Rahwan, I., Reed, C.** The argument interchange format. *Argumentation in artificial intelligence*, Rahwan I. and Simari G., Eds. Springer, 2009, pp. 383–402.
- Teruel, M., Cardellino, C., Cardellino, F., Alemany, L., Villata, S.** Increasing Argument Annotation Reproducibility by Using Inter-annotator Agreement to Improve Guidelines. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan, 2018.
- Zagorulko, Yu. A., Domanov, O. A., Sery, A. S., Sidorova, E. A., Borovikova O. I.** Analysis of the persuasiveness of argumentation in popular science texts. In: *Artificial Intelligence, Proceedings of the 18th Russian Conference RCAI*, pp. 351–367, September 2020.

Информация об авторе

Пименов Иван Сергеевич, аспирант НГУ

Information about the Authors

Ivan S. Pimenov, Postgraduate Student at Novosibirsk State University

*Статья поступила в редакцию 15.03.2023;
одобрена после рецензирования 13.04.2023; принята к публикации 14.04.2023*
*The article was submitted 15.03.2023;
approved after reviewing 13.04.2023; accepted for publication 14.04.2023*