К вопросу о формировании набора отношений для корпуса с дискурсивной разметкой текста

Е.Г.Соколова[1], С.Ю.Толдова[2]

[1]независимый исследователь,
[2]Научно-исследовательский университет «Высшая школа экономики»

minegot@rambler.ru, stoldova@hse.ru

Аннотация

Работа посвящена дискурсивной разметке корпусов. В ней анализируется состав отношений, принятых в корпусе Ru-RSTreebank. Это корпус, размеченный в рамках теории риторических структур В. Манна и С. Томпсон (Rhetoric Structure Theory, RST). При разметке корпуса был принят ряд решений относительно модификаций исходного набора отношений. В статье рассматриваются проблемы, вызванные одним из противоречий, с которым сталкиваются разработчики при создании стандартов лингвистической разметки. Это противоречие между стремлением как можно более точно отразить лингвистическую реальность, с одной стороны, и требованием обеспечить устойчивость разметки, с другой. В статье на примере дискурсивной разметки анализируются проблемы, возникающие в случае упрощения разметки для обеспечения необходимой степени согласия аннотаторов.

Ключевые слова: корпус с дискурсивной разметкой, теория риторических структур, риторические отношения

Библиографическая ссылка: Соколова Е. Г., Толдова С. Ю. К вопросу о формировании набора отношений для корпуса с дискурсивной разметкой текста // Компьютерная лингвистика и вычислительные онтологии. Выпуск 4 (Труды XXIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2020, Санкт-Петербург, 17 – 20 июня 2020 г. Сборник научных статей). — СПб: Университет ИТМО, 2020. С. 44-53. DOI: 10.17586/0000-0000-2020-4-44-53

Введение

В настоящее время отмечается новый всплеск интереса к корпусам текстов, в которых представлен слой дискурсивной разметки. Несмотря на то, что существует уже достаточно много подобных корпусов по отдельным языкам (см., например, [2; 3; 4]), активно развиваются проекты по созданию мультиязычных корпусов и универсальных стандартов дискурсивной разметки ([5; 6] и др.), вопросы, связанные с обоснованностью набора используемых дискурсивных отношений и согласованности разметки, до сих пор остаются актуальными. Повышенный интерес к уровню дискурса как в корпусной, так и в компьютерной лингвистике объясняется следующими обстоятельствами:

а) дискурсивная структура часто рассматривается в связи с другими составляющими организации текста, в частности, отношениями кореферентности (существует немало работ, в которых показано, что кореферентные отношения между языковыми выражениями не просто линейны, они «вплетены» в многомерную структуру текста, см., например, понятие фокуса внимания в [7] или теорию вен [8]);

б) информация о дискурсивной разметке оказывается актуальной при разработке современных прикладных систем в рамках таких направлений, как генерация текстов,

диалоговые системы, автоматическое реферирование и т.п. (см., например, [9]; [10]; [11]; [12]; [13]); в частности, для генерации текстов важно учитывать зону эффекта в определении Риторического отношения (РО); для анализа аргументации в дискурсе важен тип РО, для автоматического реферирования важно правильное выделение ядра (нуклеарность); нередко дискурсивно размеченные корпуса используются при построении нейросетевых моделей диалоговых систем;

г) дискурсивный анализ востребован в клинической лингвистике: исследователей интересуют особенности организации текста в речи людей с речевыми и другими нарушениями ([14]).

д) задача как можно точнее описать явления, связанные именно с организацией дискурса, выходит на первый план в когнитивно-ориентированных моделях, учитывающих взаимодействие в дискурсе всех каналов коммуникации (см. [15]; [16] и др.).

В данной статье мы анализируем возможные последствия упрощения набора РО на основе анализа разметки корпуса Ru-RSTreebank [17; 18]. В разметке этого корпуса используется набор РО, разработанный Т. Манном и С. Томпсон [1; 19], назовём его Классический набор (КН). Он в принципе применим для отображения структуры любого текста1, хотя первоначально РО создавались для генерации убеждающих текстов, а также для анализа коротких заметок из области рекламы, инструкций и политики. КН обладает следующими свойствами: а) предназначен для монологических текстов; б) предполагает определённых адресата(ов) и цель текста, которая состоит в конечном счёте в формировании у адресата(ов) намерения выполнить некоторое действие – приобрести определённый товар, поддержать политику группы лиц, лидера и др.; в) принципиально открыт для введения специализированных РО для анализа текстов разных типов. В Ru-RSTreebank представлены тексты разных типов.

Одна из проблем, с которой приходится сталкиваться при разработке стандартов и критериев разметки, это противоречие между желанием как можно более точно отразить лингвистическую реальность, которая за этим стоит, с одной стороны, и достичь высокой согласованности аннотаторов, с другой стороны. Как указывают создатели корпуса Ru-RSTreebank в [17], в связи с тем, что при первых экспериментах с разметкой не удалось достигнуть достаточно высокого показателя согласия аннотаторов, пришлось несколько перестроить первоначальную систему отношений, предложенных в работе [1].

2. Корпус Ru-RSTreebank

Корпус Ru-Rstreebank включает в себя тексты разных жанров. Первая часть — 122 текста в жанрах новостных статей и новостной аналитики, научно-популярных статей разной тематики. Вторая часть — 100 научных текстов из научной электронной библиотеки «Киберленинка» (www.cyberleninka.ru): филология и лингвистика - 50 текстов, технические и компьютерные науки – 20 текстов. Третья содержит 100 текстов блогов. Разметка корпуса проводилась несколькими аннотаторами. Корпус и материалы к нему доступны в открытом доступе [18]. В частности, в инструкции, выложенной на сайте, формулируются принципы разметки. Утверждается, что по итогам пробной разметки и экспертных обсуждений, оригинальный набор отношений, т.е. КН, был несколько изменён, в частности, сокращён за счёт объединения отношений, например, Причина-Следствие [20]. Итого на основе КН был сформирован список из 17 типов дискурсивных отношений. Для оценки согласия аннотаторов применялась мера Krippendorff’s unitized alpha, она составляет 81% для последнего измерения.

[1]Сами авторы считали, что РО не годятся для описания структуры текстов законов, контрактов и др., не имеющих собственно риторических свойств.

3. Адаптация набора риторических отношений

Учитывая тот факт, что РО в текстах принципиально не имеют формальных показателей, а лингвистические элементы, иногда сигнализирующие определённый тип РО - эксплицитные, например, «ключевые слова» и имплицитные, например, параллельные структуры, порядок слов - многофункциональны, главным источником для построения дискурсивной структуры текста являются семантические требования к РО и глубокое проникновение в смысл текста. Сложность задачи в ситуации разнотипности текстов и большого количества отношений (КН содержит около 30 РО, см. [19]) затрудняет работу разметчиков и замедляет ее, а также понижает согласованность разметки. Ниже подробно проанализируем случаи, в которых объединение некоторых отношений в одно по правилам, принятым в Ru-Rstreebank, приводит к существенному огрублению вносимой в текст при разметке информации.

3.1. Обобщение типов похожих риторических отношений

Мы обсудим набор РО, определённый для корпуса Ru-RSTreebank, см. ([17, с. 201]) по сравнению с КН. Авторы теории риторических структур (ТРС) сами группируют похожие РО: группа из пяти каузативных РО, шесть групп по два похожих РО, остальные – по одному. При настройке на корпус основания для обобщений РО различны. Мы рассмотрим решения, принятые для корпуса Ru-RSTreebank.

Каузативная группа. Вместо пяти каузативных отношений в КН: VolitionalCause, Non-VolitionalCause, VolitionalResult, Non-VolitionalResult, Purpose, используется два: Cause-Effect и Purpose, полученные в результате не различения РО по параметру Volitional / Non-Volitional и слияния, РО Result и Cause в РО Cause-Effect. Варианты Non-Volitional – действительный (объективно, в соответствии с законами действительности) и Volitional – волевой (по воле говорящего), т.е. или в физической (физ.) или в ментальной (мент.) сфере восприятия реляционного утверждения, соответственно. Параметр «действительный vs. волевой» может быть легко возвращён в разметку путём добавления к РО Cause–Effect признака со значением этого параметра.

Решение объединить отношения Cause и Result в одно РО Cause-Effect, с одной стороны, повышает согласие аннотаторов, с другой, в некоторых случаях создает структуру с ошибкой распознавания ядра и сателлита. Это существенное огрубление. При установлении несимметричных отношений сначала определяется ядро и сателлит, затем они соединяются отношением Cause-Effect. Выбор ядра предопределяет, как мы должны читать его: как Cause, таких случаев большинство, или как Effect, например:

(1) (a-b) «В феврале представители Германии, Франции, Украины и России подписали второй пакет Минских соглашений, ставших основой для перемирия и политического урегулирования конфликта. **Cause**>(c) **<Effect** *Благодаря этим договоренностям прекратились военные действия».*

Таким образом, при принятом решении определение ядра и сателлита происходит независимо ни от типа маркера отношения (например, разных подчинительных союзов), ни от его местоположения, а для исправления возможной ошибки необходимо изменить структуру составляющих.

Restatement и Multinuclear restatement

Это особый случай, когда в КН два отношения Restatement – в одном случае классифицируется как зависимость, во втором – как мультиядерное – Multinuclear Restatement (см. определение [19]). В первом случае ядро - важнее, а сателлиты усиливают внимание к ядру, риторический повтор, который, применялся ораторами античности для усиления воздействия на аудиторию, например, начало первой речи Цицерона против Катилины. Можно считать, что в наших текстах этому близок повтор при цитировании для подтверждения сказанного, например:

(2) (a)… представители Белого дома фактически одобряют и поощряют подобные взгляды. (b)"В ходе выступления госсекретаря США Джона Керри (JohnKerry) перед комитетом Палаты представителей по иностранным делам (c) стало очевидно, какой именно курс избрали в администрации Обамы.(a) **<Restatement** ((b) **Attribution**>(c)),

В наборе Ru-RSTreebank Restatement как «зависимость» отсутствует, многоядерное Restatement используется во всех случаях.

Circumstance и Background. Оба отношения задают фон, на котором воспринимается реляционное утверждение, находящееся в ядре. В первом случае «фоном» являются условия действительности, в частности, время, место, во втором – некоторая абстрактная информация, знание, позволяющее понять реляционное утверждение в ядре. В новостных и политических текстах фоном часто становится социальная обстановка событий. В наборе Ru-RSTreebank оба отношения объединены по именем Background, например, в следующем фрагменте РО Background в первом случае определяет соц. фон, во втором случае физ. фон, т.е. действительные условия, в которых сделано заявление:

(3) (Глава Белоруссии Александр Лукашенко также отметил,) (a – b) что тлеющие конфликты обострились, и возникли новые очаги нестабильности в странах Ближнего Востока, Афганистане и у границ стран ОДКБ. **Background>** *(c) "В такой ситуации требуется укрепление военной мощи и Коллективных сил оперативного реагирования",* **<Atribution** - *подчеркнул он,* **<Background** *выступая на заседании Совета ОДКБ в расширенном составе.*

При объединении РО Background и Circumstance мы пренебрегаем зоной эффекта РО, но, поскольку структуры идентичны, эта информация может быть добавлена в разметку.

Рассмотрим РО в группах, объединённых самими авторами ТРС. В них можно использовать обобщённое отношение, поскольку РО всегда можно уточнить. Так сделано в Ru-RSTreebank, но это всегда огрубление. Рассмотрим эти пары.

Interpretation и Evaluation. При разметке Ru-RSTreebank оценка считается разновидностью интерпретации. Явное огрубление в таком объединении видно при смежности фрагментов интерпретации и оценки:

(4) (a) В ряде случаев они использовались в корыстных целях для смены режимов, (b) что являлось нарушением устава ООН и основополагающих норм международного права. (c) Такое лицемерие не имеет никакого отношения к защите прав человека и продвижению идеалов свободы и демократии.

(4b) и (4c) – сателлиты к ядру (4a). Если (4b) можно отнести к интерпретации, то (4c) – это оценка. Однако в других случаях, действительно, трудно провести границу между этими двумя отношениями. Такая градуальность оценки-интерпретации подтверждается различными экспериментами по определению тональности текста. Нередко положительная или отрицательная оценка создается как раз за счет той положительной или отрицательной коннотации, которую несет предикат, который отражает интерпретацию. В данном случае предикат *нарушение* скорее обозначает факт, который расценивается в обществе как отрицательный, однако предикаты типа *усложнение*, *нарушение* в контексте *нарушение моральных норм* скорее выражают отношение автора, а не объективную реальность.

Condition и Otherwise. Определения этих отношений близки. Основное различие в том, что ядро в Otherwise – нереализованная ситуация. В отличие от Condition дискурсивная структура трёхчленна: If A then B otherwise C (см. [19]), причём вторая составляющая (then B) часто опущена, например:

(5) (a) Россия борется не только с "Исламским государством", но и вообще с джихадистскими группировками в Сирии. […] (b) Понятно, что если бы [победили террористы], (c) у нас возникло бы джихадистское государство с выходом к морю.

со структурой: (a)**<Otherwise (b) Condition>(c).** Здесь «B» в приведённой формуле представляет собой закадровый вывод «и Сирия справится с терроризмом с помощью России».

Пример (5) показывает, что в некоторых ситуациях РО Otherwise необходимо, поэтому элиминация его является огрублением.

Concession, Antithesis и Contrast. Эти три РО обобщены в наборе Ru-RSTreebank до отношения Concession. Для используемых в корпусе типов текстов его оказывается достаточно. Двухъядерное отношение Contrast рассматривалось в статье [21] наряду с введённым в набор отношением Comparison. В статье было показано, что Contrast имеет трёхчленную структуру, подобную отношению Otherwise.

(6) (a) Заключения американских спецслужб, на основании которых заговорили о работе над атомной бомбой, были не более чем догадками. **<Contrast>** *(b) Особенно нелепо эти заявления выглядят на фоне неоспоримых доказательств и свидетельств существования таких программ в Израиле, Индии, Пакистане, ЮАР и Южной Корее.*

Т.е. из (6a) следует импликатура *«(6a') доказательств иранских разработок ядерной бомбы нет»*. А уже (6a') прямо противопоставлен (6b).

Evidence и Justify. Первое довольно часто используется в текстах и при этом имеет яркие свойства, благодаря которым устанавливается с уверенностью. Второе, наоборот, свойственно в основном убеждающим текстам, а для наших текстов неактуально. Соответственно, второе имя РО в нашем наборе не используется.

3.2. Добавление типов РО

Дополнительные РО связаны с желанием отразить некоторый тип реляционного утверждения, не охватываемое КН, а также включить в дискурсивную структуру явления, отражающие коммуникативные особенности данного типа текста, например, диалог, сопоставление мнений, чат.

Compairison. Это РО добавлено для обозначения сопоставления без противопоставления. В частности, анализ научных текстов показал, что достаточно частотным приемом в научных текстах является введение и описание того или иного понятия в сравнении с уже знакомым понятием того же семантического ряда (меронимом), что позволяет читателю поместить обсуждаемый концепт в ряд других уже известных концептов и их свойств, например:

(7) Научный характер текста - в том, что он содержит художественные термины (название стилей и направлений в искусстве, материалов, техник, элементов рисунка и оборудования) и стилистически нейтральную лексику. Публицистичность искусствоведческого текста проявляется в наличии большого количества фактов, имен, дат.

В данном случае вводятся свойства искусствоведческого текста, отражающие его научный и публицистический характер, а не противопоставляются два искусствоведческих текста. В Ru-RSTreebank включение РО Comparison привело к конкуренции с РО Contrast, что потребовало дополнительного исследования (см. [21]) В результате были сформулированы некоторые методы выделения РО Comparison и обнаружена трёхчленная дискурсивная конструкция для РО Contrast.

Attribution. Разработчики Ru-Rstreebank вводят также в качестве дополнения для достижения полноты структуры для нового типа текстов – Attribution для текстов, в которых сопоставляются мнения. Оно эффективно, контаминирует с референциальной структурой, является «техническим» отношением, так как не продвигает развитие содержания текста.

Следует отметить, что этот тип отношения обладает повышенной частотой в текстах новостей по сравнению с другими жанрами. Указание источника принципиально для

текстов новостей. Во-первых, постоянная ссылка на источники в текстах позволяет (а) определить/повысить оценку сообщения как достоверного и объективного (если читатель относит источник к числу значимых и доверяет ему); (б) адекватно проинтерпретировать ядро такого отношения, если в тексте новостей содержится, напротив, не изложение факта, а значимая цитата какого-то политика; (в) соответственно, отличить объективную и субъективную информацию. Разграничение субъективной vs. объективной информации входит в круг задач анализа тональности, определение типа информации улучшает работу классификаторов по распознаванию мнений. Таким образом, такое РО может быть дополнительным признаком в системах определения тональности, основанных на традиционных подходах.

3.3. Выделение элементарных дискурсивных единиц и референциальная структура текста

Согласно стандартам выделения ЭДЕ[2] (Elementary discourse units) в TPC [1], ЭДЕ соответствуют событию. Границы ЭДЕ в прототипическом случае совпадают с границами клауз, т.е. это различные структуры с предикатом – простые предложения, придаточные, причастные обороты и некоторые другие. Но суждения, утверждения, события, т.е. фрагменты содержания текста, выступающие как ЭДЕ, могут обозначаться именными группами (ИГ). Это касается ИГ, играющих роль обстоятельств причины, цели и т.п., управляемых соответствующими предлогами (например, *для*, *из-за*). Правила предложены в [17], в соответствии с ними такие именные группы должны содержать отглагольные существительные, как, например, в ИГ «*для покупки новых тапок*». Однако существует довольно много пограничных случаев, когда ситуация выражена не отглагольным именем, а некоторым непроизводным существительным, обозначающим событие, процесс: ср. *из-за роста цен, для удобства.*

При дискурсивной разметке часто приходится мысленно восстанавливать эллипсис, например, для локативной группы в следующем примере:

(8) (a):…вариант развития событий,……..является вероятным, **<Elaboration** *(b)*: *при котором боевики ИГ устремят свой взор на Иран,***<Sequence>**(c)***:** *а за ним (боевики ИГ устремят взоры) на Кавказ и Центральную Азию … .*

Выделение из фрагмента (8a) фрагмента (8b) (рис. 1) в качестве ЭДЕ можно поставить в зависимость от наличия в этом тексте выражения, кореферентного *посягательству на Кавказ и Центральную Азию*, неоднократно возникающих в тексте.

Необходимость видеть за ИГ её подразумеваемое содержание вытекает из самой задачи дискурсивного анализа, цель которого представить именно всё содержание в виде одной согласованной и законченной структуры.

Заключение

В статье рассмотрены различные последствия упрощения и модификации РСТ разметки, принятые при разметке корпуса Ru-RSTreebank. Описан результат настройки Классического набора риторических отношений на нужды данного корпуса. В целом обобщение и отказ от некоторых отношений вполне обоснован. Однако выделено два случая нежелательного отказа от РО. Это РО Result или Effect и Condition или Othewise. В первом случае желательно вернуться к двум отношениям Cause и Effect при том, что выбор ядра обязательно предшествует установлению отношения. Во втором Othewise

[2] В российской традиции сокращение «ЭДЕ»: «Элементарная дискурсная единица» иногда ассоциируется с анализом устного дискурса, для которого было впервые использовано. На самом деле это – общеупотребительный термин – Elementary discourse unit (EDU), определяемый для любого текстового или звучащего корпуса его составителями.

нечем заменить. Следует ещё раз рассмотреть функции добавленного отношения Comparison и возможность его обобщения с PO List. Полученный опыт будет полезен для дальнейшей разметки корпуса Ru-RSTreebank, а также для настройки других русскоязычных корпусов и для сравнения с иноязычными дискурсивно размеченными корпусами, поскольку унификация дискурсивной разметки важна для сравнительного исследования структур текстов разных культур.

Авторы выражают благодарность Российскому фонду фундаментальных исследований (гранты №17-29- 86 07033) за финансовую поддержку данной работы.

Литература

[1] Mann W.C., Thompson S.A. Rhetorical structure theory: Toward a functional theory of text organization // Text-Interdisciplinary Journal for the Study of Discourse. 1988. №8(3). P. 243–281.
[2] Cunha I., Torres-Moreno J., Sierra G. On the development of the RST Spanish treebank // Language Resource Evaluation. 2015. Vol. 49, № 2. P. 263309. DOI: https:// doi.org/10.1007/s10579-014-9271-6.
[3] Al-Saif A., Markert K. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic // Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010). 2010. P. 2046–2053.
[4] Carlson L., Marcu D., Okurowski M.E. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory // Current directions in discourse and dialogue, Kluwer Academic Publishers. 2003. P. 85–112.
[5] Iruskieta M., Cunha I., Taboada M. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora // Lang. Resour. 2015. Vol. 49, 2. P. 263–309. DOI: https://doi.org/10.1007/s10579-014-9271-6.
[6] Zeyrek D., Mendes A., Grishina Y., Kurfalı M., Gibbon S., Ogrodniczuk M. TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style // Lang Resources & Evaluation. 2020. № 54. P. 587–613. DOI: 10.1007/s10579-019-09445-9.
[7] Grosz B.J., Sidner C.L. Attention, intentions, and the structure of discourse Comput. 1986. № 12, 3. P. 175–204.
[8] Cristea D, Ide N, Romary L. Veins theory: A model of global discourse cohesion and coherence // Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. 1998. Vol. 1. P. 281-285.
[9] Hirao T., Yoshida Y., Nishino M., Yasuda N., Nagata M. Single-document summarizationas a tree knapsack problem // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013. P. 1515–1520.
[10] Somasundaran S. Discourse-level relations for Opinion Analysis. PhD thesis, University of Pittsburgh. 2010.
[11] Chai J.Y., Jin R. Discourse structure for context question answering // Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL, 2004. P. 23-30.
[12] Qin B., Tang D., Geng X., Ning D., Liu J., Liu T. A planning based framework for essay generation. arXiv preprint arXiv:1512.05919, 2015.
[13] Afantenos S., Kow E., Asher N., Perret J. Discourse parsing for multi-party chat dialogues // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015. P. 928-937.
[14] Кибрик А., Подлесская В. Рассказы о сновидениях: Корпусное исследование устного русского дискурса // Litres, 2017. 733 с.
[15] Abuczki A., Esfandiari B. G. An overview of multimodal corpora, annotation tools and schemes // Argumentum. 2013. Vol. 9. P. 86–98.

[16] Кибрик А.А., Федорова, О.В., Подлесская, В.И. Мультиканальные корпуса: вчера, сегодня, завтра // А. Б. Безбородов (ред.) Гуманитарные чтения РГГУ. М: РГГУ. 2017. С. 499–511.
[17] Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A, Nikiforova S., Pavlova I., Shelepov A. Towards building a discourse-annotated corpus of Russian // Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue. 2017. № 16. P. 23.
[18] Ru-RSTreebank. Русскоязычный дискурсивный корпус. URL: https://www.rstreebank.ru (дата обращения: 22.05.2020).
[19] Rhetorical Structure Theory (RST). URL: https://www.sfu.ca/rst/ (дата обращения: 22.05.2020).
[20] Toldova S., Pisarevskaya D., Kobozeva M., Vasilyeva M. The cues for rhetorical relations in Russian:“Cause-Effect” relation in Russian rhetorical structure treebank // Comput. Linguist. Intellect. Technol. 2018. Vol. 17, № 24. P. 748-761.
[21] Toldova S., Davydova T., Kobozeva M., Pisarevskaya D. Contrast and Comparison Relations in RST Framework: the Case of Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”. 2019. P. 714-727.

On the Modification of Standard Discourse Relations Set for the Annotation of Ru-RSTreebank

E. Sokolova[1], S. Toldova[2]

[1] independent researcher, [2]National Research University “Higher School of Economics”

The work discusses the problem of discourse annotation and the consequences of the relations set simplification for the sake of higher interannotator agreement. One of the theoretical approaches to discourse structure representation is the Rhetoric Structure theory by William Mann and Sandra Thompson [1]. There is a set of rhetoric relations between discourse units that were suggested within this theory. Modern discourse treebanks, including the discourse annotated treebank for Russian (Ru-RSTreebank), exploit this set of relations. Considering the conceptual rather than operational basis of relations definitions and the lack of strict grammatical tests for specifying relation types, the developers of these treebanks face the problem of low interannotator agreement. One of the ways to overcome this difficulty is to simplify the annotation scheme, e.g. to unite confusing relations in one relation. The developers of the Ru-RSTreebank chose this option. We discuss the pitfalls of this approach. We suggest the detailed analysis of the consequences for these decisions exemplified with the data from Ru-RSTreebank. One of the “tree-constructing” issues is the direction of a relation between two nodes. Thus, according to [17], the annotators often confuses the relation direction between two discourse units, one expressing the cause and the other expressing the result. The two different directionsmatch the two different relations (‘Cause’ and ‘Result’) in the standard set of rhetoric relations. The developers decided to unite these two relations and to suggest the only one possible direction for them. According to our analysis, the effect of this operation is that the real nuclear discourse unit that should form the basis of the text changes its status to a satellite. Consequently, the logical structure of the text skeleton is distorted. In our work, we suggest the analysis of some other cases. Besides, one of the drawbacks is that the annotation scheme loses its universality that makes it less comparable to the treebanks for other languages.

Keywords: discourse annotation, rhetoric structure theory, rhetoric relations
Reference for citation: Sokolova E.G., Toldova S. On the modification of standard discourse relations set for the annotation of Ru-RSTreebank // Computational linguistics and computational

ontologies.Vol. 4 (Proceedings of the XXIII International Joint Scientific Conference «Internet and Modern Society», IMS-2020, St. Petersburg, June 17-20, 2020). - St. Petersburg: ITMO University, 2020. P. 44 – 53. DOI: 10.17586/0000-0000-2020-4-44-53

References

[1] Mann W.C., Thompson S.A. Rhetorical structure theory: Toward a functional theory of text organization // Text-Interdisciplinary Journal for the Study of Discourse. 1988. №8(3). P. 243–281.
[2] Cunha I., Torres-Moreno J., Sierra G. On the development of the RST Spanish treebank // Language Resource Evaluation. 2015. Vol. 49, № 2. P. 263–309. DOI: https:// doi.org/10.1007/s10579-014-9271-6.
[3] Al-Saif A., Markert K. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic // Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010). 2010. P. 2046–2053.
[4] Carlson L., Marcu D., Okurowski M.E. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory // Current directions in discourse and dialogue, Kluwer Academic Publishers. 2003. P. 85–112.
[5] Iruskieta M., Cunha I., Taboada M. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora // Lang. Resour. 2015. Vol. 49, 2. P. 263–309. DOI: https://doi.org/10.1007/s10579-014-9271-6.
[6] Zeyrek D., Mendes A., Grishina Y., Kurfalı M., Gibbon S., Ogrodniczuk M. TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style // Lang Resources & Evaluation. 2020. № 54. P. 587–613. DOI: 10.1007/s10579-019-09445-9.
[7] Grosz B.J., Sidner C.L. Attention, intentions, and the structure of discourse // Comput. 1986. № 12, 3. P. 175–204.
[8] Cristea D, Ide N, Romary L. Veins theory: A model of global discourse cohesion and coherence // Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. 1998. Vol. 1. P. 281-285.
[9] Hirao T., Yoshida Y., Nishino M., Yasuda N., Nagata M. Single-document summarizationas a tree knapsack problem // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013. P. 1515–1520.
[10] Somasundaran S. Discourse-level relations for Opinion Analysis. PhD thesis, University of Pittsburgh. 2010.
[11] Chai J.Y., Jin R. Discourse structure for context question answering // Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL, 2004. P. 23-30.
[12] Qin B., Tang D., Geng X., Ning D., Liu J., Liu T. A planning based framework for essay generation. arXiv preprint arXiv:1512.05919, 2015.
[13] Afantenos S., Kow E., Asher N., Perret J. Discourse parsing for multi-party chat dialogues // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015. P. 928-937.
[14] Kibrik A., Podlesskaya V. Rasskazy o snovideniyah: Korpusnoe issledovanie ustnogo russkogo diskursa // Litres, 2017. 733 s. [In Russian].
[15] Abuczki A., Esfandiari B. G. An overview of multimodal corpora, annotation tools and schemes // Argumentum. 2013. Vol. 9. P. 86–98.
[16] Kibrik A. A., Fedorova, O. V., Podlesskaya, V. I. Mul'tikanal'nye korpusa: vchera, segodnya, zavtra // A. B. Bezborodov (red.) Gumanitarnye chteniya RGGU. M: RGGU. 2017. S. 499–511.
[17] Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A. Towards building a discourse-annotated corpus of Russian // Computational

Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue. 2017. № 16. P. 23.

[18] Ru-RSTreebank. Russkoyazychnyj diskursivnyj korpus. URL: https://www.rstreebank.ru (data obrashcheniya: 22.05.2020). [In Russian].

[19] Rhetorical Structure Theory (RST). URL: https://www.sfu.ca/rst/ (data obrashcheniya: 22.05.2020). [In Russian].

[20] Toldova S., Pisarevskaya D., Kobozeva M., Vasilyeva M. The cues for rhetorical relations in Russian:“Cause-Effect” relation in Russian rhetorical structure treebank // Comput. Linguist. Intellect. Technol. 2018. Vol. 17, № 24. P. 748-761.

[21] Toldova S., Davydova T., Kobozeva M., Pisarevskaya D. Contrast and Comparison Relations in RST Framework: the Case of Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”. 2019. P. 714-727.