ИЗДАТЕЛЬСТВО ГРАМОТА

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2024. Том 17. Выпуск 9 | 2024. Volume 17. Issue 9
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Согласованность читателей при разметке элементов текстовых миров в корпусе TextWorlds

Михалькова Е. В.

Аннотация. С точки зрения теории текстовых миров (Text World Theory), в нарративах содержатся элементы (указания времени, места, персонажей и др.), которые можно выявлять автоматически и сравнивать на их основе версии событий и схожие сюжеты. Мы разметили корпус сказок и коротких рассказов TextWorlds и обнаружили, что читатели не всегда согласны в том, указывает ли то или иное слово на персонажа, время или место действия. Цель исследования – установить степень согласованности читателей относительно положения этих нарративных категорий в тексте. Практическая задача исследования заключается в оценке надежности разметки, которая будет использоваться для обучения алгоритмов выявлению текстовых миров автоматически. Научная новизна заключается в том, что мы изучаем именно степень согласованности, в то время как в других работах согласованность рассматривается как нечто само собой разумеющееся, и если читатели не согласны друг с другом, то это воспринимается как ошибка одного из читателей или процедуры разметки. В статье мы представляем результаты двух метрик согласованности экспертов: процент согласия и альфа Криппендорфа. Полученные результаты для этих метрик показывают, что согласованность относительно разных элементов варьируется в зависимости от произведения и иногда достигает среднего уровня, достаточного для того, чтобы говорить о надежности разметки.

EN

Inter-rater agreement in annotating text world elements in the TextWorlds corpus

E. V. Mikhalkova

Abstract. From the perspective of Text World Theory, narratives contain elements (indications of time, place, characters, etc.) that can be automatically identified and compared to establish versions of events and similar plots based on these elements. We have annotated a corpus of fairy tales and short stories, TextWorlds, and discovered that raters do not always agree on whether a particular word refers to a character, time, or place of action. The aim of the research is to determine the degree of inter-rater agreement regarding the position of these narrative categories in the text. The practical task of the research is to assess the reliability of the annotation that will be used to train algorithms for automatically identifying text worlds. The scientific novelty lies in the fact that we are specifically studying the degree of agreement, whereas in other works, agreement is taken for granted, and if raters disagree with each other, it is perceived as an error by one of the raters or the annotation procedure. In this paper, we present the results of two expert agreement metrics: percent agreement and Krippendorff's alpha. The obtained results for these metrics show that agreement regarding different elements varies depending on the work and sometimes reaches an average level, sufficient to speak of the reliability of the annotation.

Введение

Машинное восприятие текста (“machine reading comprehension” (Sang, Mou, Li et al., 2022)) – это развивающийся раздел компьютерной лингвистики, который моделирует чтение таким образом, чтобы машинная интерпретация текстов напоминала человеческую, но облегчала при этом обработку больших массивов текстовых данных. Это нужно, например, для выявления сходств в сюжетных линиях и переводах, классификации сюжетов, моделирования игровых процессов, а также сопоставления свидетельских показаний (Евсеев, Кох, Михалькова, 2023).

Теория текстовых миров (Text World Theory), которая появилась в 1980-е гг., не так давно стала активно использоваться в компьютерной лингвистике для моделирования машинного восприятия повествования

Научная статья (original research article) | https://doi.org/10.30853/phil20240451

через когнитивные конструкты под названием «текстовые миры». Текстовый мир – это репрезентация авторского и/или читательского видения событий. Предтечей теории считается М. М. Бахтин, который ввел в литературоведение термин «хронотоп» – «существенную взаимосвязь временных и пространственных отношений, художественно освоенных в литературе» (1975, с. 234).

Мы создали корпус из нарративных произведений, в которых явно прослеживаются истории, последовательности событий (фольклорные сказки, авторские рассказы и отрывки повествований из повестей), и разметили в них признаки текстовых миров (Mikhalkova, Protasov, Drozdova et al., 2019; Mikhalkova, Protasov, Gavin et al., 2020). При этом мы обнаружили, что к разметке текстовых миров трудно применимы критерии согласованности экспертов, что неудивительно, т. к. многие художественные тексты намеренно создаются так, чтобы возникла неоднозначность и остался простор для интерпретации. Наш эксперимент с разметкой показал, что читатели не всегда согласны в том, указывают ли конкретные слова даже на такие базовые категории, как время, место действия и персонажи.

Мы ставим перед собой следующие задачи: измерить согласованность читателей метриками «процент согласия» и «альфа Криппендорфа»; рассмотреть некоторые типичные случаи несогласия; определить особенности нарратива и нашей разметки, на которые указывает несогласие читателей.

Актуальность нашего исследования обусловлена сложившимися отличиями во взглядах на правильность определения нарративных категорий в художественном тексте. С одной стороны, эксперты-филологи и обычные читатели предлагают разнообразные интерпретации текста и подчеркивают, что художественная экспрессия должна приводить к разным трактовкам. С другой стороны, компьютерные лингвисты считают: высокая согласованность экспертов при разметке текста – это то, чего нужно достичь, что нужно увеличивать. При этом делается ставка на то, что в разметке есть только один правильный вариант. Это приводит к тому, что при аннотировании уменьшается разнообразие интерпретаций, пропадают особенности самого опыта чтения. Поэтому разметка корпусов, в особенности художественной речи, сегодня находится в своего рода кризисе, о чем ведутся оживленные дискуссии (Beck, Booth, El-Assady et al., 2020; Uma, Fornaciari, Dumitrache et al., 2021; Peng, Sun, Loftus et al., 2024; Weber-Genzel, Peng, De Marneffe et al., 2024).

Выбор методов исследования обусловлен целью и совокупностью поставленных задач. Для исследования согласованности читателей был выбран количественный метод межэкспертной надежности, представленный метриками «процент согласия» и «альфа Криппендорфа». Кроме этого, сравнительно-сопоставительный метод был использован для выявления сходств и различий в типичных случаях несогласия, а также для описания особенностей нарратива и разметки.

Материалом исследования послужила коллекция размеченных текстов:

- Михалькова Е. В., Протасов Т. А., Гэвин П. И., Башмакова А. Ю., Дроздова А. О., Жмыхов А. Г. Текстовый корпус TextWorlds. https://github.com/evrog/TextWorlds.

Теоретическую базу работы составили труды теоретиков в области текстовых миров: П. Верта (Werth, 1993; 1994; 1999), П. Стоквелла (Stockwell, 2020), Дж. Гэвинз (Gavins, 2007) и др. (Bell, Ryan, 2019; Gibbons, Whiteley, 2020; Raghunath, 2020; Sirinarang, Wijitsopon, 2021), исследования по выявлению указаний времени, места, действующих лиц в тексте (Detkova, Novitskiy, Petrova et al., 2020; Ho, Lugea, McIntyre et al., 2018; Honnibal, Montani, Van Landeghem et al., 2020; Hu, Mao, McKenzie, 2019; Srivatsa, Srinivasa, 2018), а также результаты экспериментов по разметке текстов (Ho, Lugea, McIntyre et al., 2019; Wang, Ho, Xu et al., 2006; Евсеев, Кох, Михалькова, 2023). Сравнение метрик оценки согласованности представлено работами Дж. Коэна (Cohen, 1960), К. Криппендорфа и А. Хейеса (Hayes, Krippendorff, 2007). Также ранее мы предложили методологию разметки элементов текстовых миров и опробовали ее на художественных текстах, которые находятся в открытом доступе (Mikhalkova, Protasov, Drozdova et al., 2019; Mikhalkova, Protasov, Gavin et al., 2020).

Практическая значимость работы: материалы исследования могут быть использованы для изучения опыта читательской деятельности, интерпретации художественных текстов, сопоставления сюжетов и выявления в них элементов хронотопа, включая разработку инструментов автоматизации.

Обсуждение и результаты

Корпус текстовых миров

Автоматическое извлечение элементов текстовых миров представляет интерес для компьютерной лингвистики, а именно той ее части, которая занимается обработкой нарративов (в широком смысле – любых историй, описаний событий), включая как распознавание, так и генерацию. Методы машинного обучения, которые повсеместно применяются в компьютерной лингвистике сегодня, требуют разработки корпусов данных для тренировки и тестирования алгоритмов.

В открытой части нашего корпуса TextWorlds на данный момент находится 20 текстов (см. Табл. 1). Разметка текстов производится в формате xml: токенам, составляющим текст, присваиваются теги. Токен – это слово, пунктуационный символ или любая другая единица языка, которая является минимальной при оценке; в нашем случае токенами становятся обычно отдельные слова и пунктуационные символы. К токену добавляется открывающий и закрывающий тег: «В развесистых <px>платанах</px> пели <cg>соловьи</cg>, в лавровых <px>кустах</px> без умолку звенели <cg>цикады</cg>…» («Дочь Севера»). При этом несколько токенов могут получить один и тот же тег. Разметка производится при помощи программы Sublime Text Editor.

Таблица 1. *Распределение тегов в размеченных текстах*

№	Название произведения, автор (в скобках)	Всего текстов		Всего присвоено тегов								
		Переходы	Элементы	ts	ps	ms	t	p	px	c	cx	cg
1	Визит (Саша Черный)	0	1	0	0	0	0	33	0	178	0	0
2	Дочь Севера (В. М. Дорошевич)	3	5	32	86	13	38	163	58	292	34	71
3	Зеленая лампа (Александр Грин)	0	2	0	0	0	93	53	38	349	14	42
4	Золушка (Е. Шварц, отрывки)	1	1	5	19	11	0	59	7	518	6	0
5	Золушка (Ш. Перро, пер. Т. Габбе)	1	1	37	19	13	0	53	8	373	6	0
6	Лисичка-сестричка и волк (А. Н. Афанасьев)	0	6	0	0	0	15	114	28	579	9	14
7	Лягушка-путешественница (В. М. Гаршин)	1	1	19	27	5	19	46	9	190	5	12
8	Мирная война (Саша Черный)	1	1	1	11	6	11	21	54	78	29	0
9	Они выросли рядом (К. А. Морозова)	0	1	0	0	0	13	30	12	84	6	52
10	Попугай: Майкина сказка (Н. Г. Гарин-Михайловский)	5	1	140	247	30	48	65	8	181	1	48
11	Потец (А. И. Введенский)	1	1	3	27	6	0	16	9	198	0	1
12	Похождения бравого солдата Швейка (гл. 1, ч. 2, пер. П. Богатырева)	1	1	220	112	0	0	96	73	613	72	90
13	Про Иванушку-дурачка (М. Горький)	2	2	10	38	7	4	64	17	290	9	88
14	Сказка о сером волке (из сборника «Старая погудка на новый лад: русская сказка в изданиях конца XVIII века», автор не указан)	1	4	12	34	6	33	175	23	715	5	20
15	Соломинка, уголек и боб (пер. К. Азадовского)	1	1	0	0	0	0	25	5	51	0	0
16	Соломинка, уголек и боб (пер. А. Введенского)	1	1	0	0	0	0	31	3	57	0	0
17	Соломинка, уголек и боб (пер. Г. Петникова)	1	1	0	0	0	0	28	4	50	0	0
18	Умный работник: русская народная сказка (в сборнике «Сказки народов мира», автор не указан)	0	1	0	0	0	19	36	17	133	2	22
19	The Gift of the Magi (O. Henry)	4	4	415	36	0	132	128	15	1023	43	79
20	The Good Soldier Schweik (гл. 1, ч. 2, пер. Paul Selver)	1	1	113	83	1	0	61	10	452	29	54
Всего		25	37	1007	739	98	425	1297	398	6404	270	593
Процент от общей суммы				55	40	5	5	14	4	68	3	6

Для корпуса были выбраны тексты из открытого доступа, чтобы соблюсти авторское право на размещение в репозитории. Большинство текстов представляют собой сказки и короткие рассказы, что, во-первых, обеспечивает наличие нарратива, во-вторых, не позволяет аннотатору сильно устать во время чтения или потерять счет персонажам. Однако в корпусе есть и исключения: отрывок главы «Злоключения Швейка в поезде» из книги «Похождения бравого солдата Швейка» представлен на русском и английском языках для сравнения переводных версий. Также размеченный отрывок из пьесы Е. Шварца «Золушка» демонстрирует сюжетные отличия от сказки Шарля Перро, написанной как повествование от третьего лица.

Ученые по-разному смотрят на мирообразующие элементы, их типы и иерархию, но для читателя, аннотирующего корпус, методология разметки должна быть однозначной и интуитивно-понятной. Следовательно, в ней не должно быть специфической терминологии, например «дейктическое пространство» или «буломатический подмир», а также набор тегов должен быть небольшим, чтобы аннотатор «не ленился» подобрать тег, который наиболее точно отражает его читательский опыт. На основе теоретической литературы о текстовых мирах и предыдущих экспериментов по разметке текстов на английском языке мы создали модель разметки и инструкцию для аннотаторов.

Мы остановились на разметке следующих элементов:

1) переходы между текстовыми мирами;
2) указатели времени;
3) указатели места;
4) персонажи.

Разметка каждого текста производится в два слоя: в одно прочтение размечаются переходы, затем все остальные элементы.

Переходы сигнализируют о смене текстового мира, следовательно, сам текстовый мир размечать не нужно. Достаточно указать, что в данном месте текста произошел переход. Аннотация переходов происходит за одно прочтение текста, т. е. во время разметки больше никаких элементов искать не нужно. Для переходов необходимо указать тип: смена времени (тег <ts>) или локации (<ps>). Если разметчик не уверен в типе, то может поставить тег неопределенного перехода <ms>.

Указатели времени и места и персонажи также размечаются за одно прочтение. Эти три элемента обычно не накладываются друг на друга и хорошо различаются в тексте интуитивно. Персонажи и указатели места могут быть в тексте постоянными, например главные герои или место, где часто происходят события сюжета. В таком случае к тегу добавляется номер, например <p1> или <c1>, который присваивается элементам при первом их упоминании в порядке появления. Если персонаж или локация единичные, то к тегу добавляется указатель “x” (<px>, <cx>). Для групп персонажей используется указатель “g” (<cg>).

Таблица 1 содержит сведения о текстах, количестве аннотаторов и размеченных элементов в нашем корпусе.

Анализ согласованности читателей в разметке элементов текстовых миров

В компьютерной лингвистике принято оценивать согласованность аннотаторов в разметке, т. е. определять при помощи метрики, насколько совпали теги, расставленные разными людьми. Случаи, когда оценки экспертов могут не совпадать, но быть приблизительно одинаковыми, называются “inter-rater reliability” (межэкспертная надежность) и охватывают все рейтинговые способы оценивания. Например, шкала от 1 до 5 предполагает, что чем ниже оценка, тем хуже выполнено задание, и наоборот. В таком случае если разметчики поставили 4 или 5, то их оценки очень близки, а если один разметчик ставит 1, а другой – 5, то их оценки слишком разнятся, чтобы говорить о согласованности. Если оценивается именно точное совпадение тегов, то измеряется “inter-rater agreement” (согласие экспертов), как в нашем случае. Для него используются такие метрики, как:

1) *процент оценок*, которые совпали у всех разметчиков, в отношении к количеству всех размеченных случаев (“proportion or percentage of agreement (P)” (Tinsley, Weiss, 1975)), также называемый «процент согласия» (“percent agreement”). Эта метрика применяется редко, т. к. не ясно, как оценить полученный процент применительно к конкретному случаю;

2) *каппа Коэна* (Cohen, 1960) сравнивает разметку с ситуацией, когда теги были поставлены случайно, а ее более поздний вариант – взвешенная каппа (Cohen, 1968) – адаптирует метрику для рейтинговых оценок. Основное ограничение этой метрики заключается в том, что она рассчитывается только для двух разметчиков. Если их больше, берется среднее от попарного сравнения для каждых двух разметчиков. Если каппа равна 1, то это свидетельство полного согласия. Однако не существует точного представления о том, какой показатель метрики ниже 1 считать достаточным, чтобы можно было говорить о согласованной разметке (Bakeman, Quera, McArthur et al., 1997);

3) *пи Скотта* «корректирует процент согласия на количество категорий в коде и частоты, с которой каждая категория использована» (Олейник, Попова, Кирдина и др., 2014, с. 101). *Каппа Фляйса* (Fleiss, 1971) расширяет ее применение для более чем двух разметчиков. У этих метрик также трудности с интерпретацией. J. R. Landis, G. G. Koch (1977) приводят схему оценки согласия при помощи данной метрики, однако K. L. Gwet (2014) считает, что от подобных схем может быть больше вреда, чем пользы;

4) *альфа Криппендорфа* (Krippendorff, 2018; Hayes, Krippendorff, 2007) – это метрика, в основе которой заложено представление об истинном распределении категорий. Для этого оно сравнивается с ситуацией, когда разметчики расставляют теги в случайном порядке. Метрику можно использовать для двух и более разметчиков, а также если есть пропуски в разметке, т. е. если кто-то пропустил элемент. Альфа становится менее точной с ростом числа категорий (Олейник, Попова, Кирдина и др., 2014) и может принимать отрицательные значения, если согласие экспертов меньше, чем случайное. Тогда можно говорить, что точки зрения экспертов расходятся полярно.

Разный читательский опыт выражается во множестве интерпретаций, которые находят отражение как в экспертной филологической литературе, так и в обычном обсуждении прочитанного. Тем не менее мы можем ожидать, что у этой неоднозначности есть границы. В интерпретации мы понимаем, что речь идет о конкретном литературном тексте. А если интерпретация заходит слишком далеко, то большинство читателей с ней вряд ли согласится. Следовательно, разметку текстовых миров можно сравнить с голосованием экспертов, которое лучше всего моделируется такой метрикой, как процент согласия.

С другой стороны, если рассматривать читателя как «расшифровывателя» текста, интерпретатора, то его косвенной целью становится выявление всех истинных категорий текста (в нашем исследовании – элементов текстовых миров). В таком случае альфа Криппендорфа хорошо моделирует эту ситуацию.

В нашем корпусе TextWorlds несколько текстов были размечены двумя и более читателями (см. Табл. 1, «Всего аннотировано текстов»). В Таблице 2 приведены метрики «процент согласия» и «альфа Криппендорфа» для трех текстов, в которых было наибольшее количество разметчиков. Например, элементы в сказке «Лисичка-сестричка и волк» разметили шесть читателей. Метрики приводятся сначала для расчета согласия у всех токенов в тексте, а затем только у тех, к которым поставил тег хотя бы один разметчик. Так мы можем оценить согласованность внутри тех отрезков текста, которые хотя бы одному читателю показались

маркированными, т. е. похожими на элемент текстового мира. Маркированных отрезков меньше, чем не маркированных, следовательно, мы также снижаем эффект от этого дисбаланса.

Таблица 2. *Согласованность читателей в разметке элементов текстовых миров*

Название	Кол-во разметчиков	Всего тегов	Тег	Всего тегов	Процент согласия, %		Альфа Криппендорфа	
					Все токены	Маркир-е	Все токены	Маркир-е
Дочь Севера	3	3052	ts	132	96	44	0.49	-0.12
			ps	346	83	36	0.14	-0.33
			ms	47	97	33	0.11	-0.4
	5	5095	t	79	98	50	0.28	-0.08
			p	381	92	53	0.28	-0.04
			px	127	96	54	0.11	-0.17
			c	366	95	61	0.54	0.24
			cx	46	98	59	0.05	-0.21
			cg	83	98	56	0.32	-0.01
Попугай: Майкина сказка	5	6967	ts	317	94	53	0.33	-0.03
			ps	552	93	58	0.53	0.14
			ms	118	97	60	0.07	-0.25
Лисичка-сестричка и волк	6	3938	t	28	99	50	0.25	-0.1
			p	236	98	79	0.82	0.54
			px	87	97	53	0.38	0.04
			c	703	94	77	0.8	0.42
			cx	12	99	54	0.2	-0.09
			cg	14	99	63	0.44	0.17

Метрики были рассчитаны при помощи модуля metrics.agreement библиотеки NLTK (Bird, Klein, Loper, 2009). Метрика процента согласия посчитана при помощи метода avg_Ao и переведена в процент, хотя в документации она указана как среднее согласие по всем экспертам (https://www.nltk.org/api/nltk.metrics.agreement.html). В расчете процента согласия для элементов сказки «Дочь Севера» мы исключили токены, у которых было два и более тега, т. к. в NLTK реализация алгоритма не позволяет обрабатывать такие случаи. Всего таких токенов было 891.

Самые высокие показатели – у процента согласия для всех токенов: за исключением одного случая, он выше 90%. Большая часть этого результата обусловлена токенами, у которых нет тега, т. е. читатели в целом согласны, что основной массив текста не содержит элементов текстовых миров. При этом среди маркированных токенов самый высокий процент согласия наблюдается в сказке «Лисичка-сестричка и волк»: у тега «место» (p) – 79% и у тега «персонаж» (c) – 77%. При этом теги времени t и ts показывают самый низкий результат, кроме одного случая. Несогласованность экспертов может быть следствием того, что в данных текстах время выражено, вербализовано менее определенно, чем место и персонажи.

Альфа Криппендорфа согласуется с тем, что показывает процент согласия для маркированных токенов. У нее также высокие значения для тегов места и персонажей в сказке «Лисичка-сестричка и волк», а также для тега «временной переход» (ts) и «персонаж» (c) в сказке «Дочь Севера» и для тега «пространственный переход» (ps) в рассказе «Попугай: Майкина сказка». Из этого можно сделать вывод, что для разных произведений разные элементы текстовых миров могут быть более явно выражены и, следовательно, приводят к меньшим расхождениям в интерпретации. Это можно объяснить и жанровой природой, например, в пересказах фольклорных сказок, близких к оригиналу, персонажи и локации обычно задаются одними и теми же словами, без синонимов и красочных описаний. Замены осуществляются в основном через использование местоимений, например, «она» вместо «лисичка», что вряд ли собьет читателя с толку. При этом маркеры времени почти отсутствуют, т. к. само изложение предполагает историческое следование событий, одного за другим. Здесь нет, например, запутанности во времени, характерной для фантастики, или недосказанности, как в детективах.

При разработке корпуса у нас была гипотеза, что деление на текстовые миры и определение переходов между ними будет менее согласованным и, следовательно, более субъективным, чем определение других элементов. Смену текстового мира можно сравнить со сменой сцен в спектакле или фильме. Существенное изменение в пространстве и/или времени, а, возможно, и в составе присутствующих в сцене персонажей как будто запускает новую сцену. Но что считать существенным изменением? Здесь мы и обращаемся к читателям, чтобы они разметили переключатели между текстовыми мирами и стали как бы режиссерами своей интерпретации нарратива. В Таблице 2 видно, что разметка переходов даже в самом лучшем случае получила более низкие оценки согласованности (альфа = 0.49, 0.53) по сравнению с лучшими результатами других элементов (альфа = 0.54, 0.8, 0.82). Однако разница не очень значительная. Мы предполагаем, что переходы зависят от элементов хронотопа, а не наоборот. Поэтому несогласованность во вторых приводит к еще большей несогласованности в первых. Следовательно, при большей согласованности в разметке указателей времени, места и персонажей мы, предположительно, получим более высокую согласованность в разметке переходов.

Типичные случаи расхождений в разметке

Наш корпус демонстрирует, что обычно большинство читателей согласно с определением элемента текстового мира. Однако есть и ситуации, где расхождения можно назвать типичными. Рассмотрим их на примере рассказа «Мы вам все припомним» (“Total Recall”) Филипа К. Дика в переводе Владимира Баканова и Марины Осиповой. В эксперименте с разметкой этого рассказа, который мы не можем выложить в открытый доступ в связи с авторским правом, проявили себя многие особенности, которые потом появились и в других текстах.

Читатели склонны по-разному оценивать роль упоминаемых персонажей и локаций в художественном тексте, чтобы отметить их как самостоятельные элементы (т. е. поставить тег с номером, а не *px* или *cx*). Будучи читателями, мы ожидаем, что в повествовательном тексте будет больше мест и персонажей, которые останутся с нами на протяжении всей книги. Иначе нам придется запомнить много случайных элементов и потом разочароваться от того, что они мало повлияли на рассказанную историю. Произведения, где поставлен такой эксперимент, это, пожалуй, «Улисс» и «Поминки по Финнегану» Дж. Джойса. В нашем эксперименте тексты выбраны так, что хронотоп в них не «распадается» (можно сказать, что он является каноническим), поэтому несогласие читателей объясняется особенностями элементов.

Рассмотрим самое начало рассказа «Мы вам все припомним»: «Он проснулся и… захотел полететь на Марс. Его долины… что бы он ощущал, если бы бродил по ним? Величие, бесконечное величие, мечта все больше охватывала его по мере того, как он просыпался. Он чувствовал обволакивающее присутствие другого мира, который могли увидеть только правительственные чиновники да высокопоставленные особы. А простой клерк, как он? Никогда». Здесь разметчики по-разному отнесли «долины» и «другой мир» к категории «место». Кто-то посчитал, что все они относятся к месту «Марс», а кто-то счел их отдельными локусами. Основанием для включения долин в Марс можно считать отношение «часть – целое» между этими сущностями. Однако нам еще не ясно, будет ли действие рассказа происходить на Марсе или даже, более частным образом, в его долинах. Также не ясно, находятся ли Марс и другой мир в отношениях эквивалентности. Возможно, другой мир для главного героя шире, чем только планета Марс. Т. к. по идее само чтение и разметка происходят линейно, то эти предположения сказываются на различиях в расстановке тегов. Тем не менее мы не можем гарантировать эту линейность. Мы ожидаем от читателя, что он может по собственной воле читать текст не линейно, забегать вперед и возвращаться назад. Такое же поведение мы можем ожидать от разметчиков текста; в нашем эксперименте их движение по тексту не фиксируется. Сама ситуация разметки может провоцировать такое забегание тем, что разметчик сразу хочет поставить правильный (!) тег. Следовательно, вполне возможно, что кто-то из разметчиков отредактировал тег после того, например, как забежал вперед и проверил, будет ли действие рассказа происходить на Марсе (т. е. поставил тег с номером), а кто-то поставил для начала тег без номера (что означает «случайная локация, упомянутая вскользь»), а потом, когда Марс снова появлялся в рассказе, ставил тег уже с номером.

Отношения «часть – целое» вызывают сомнения и в случае более маленьких объектов, например письменного стола. В примере «Он порылся в ящике стола» стол и ящик были размечены и как отдельные локации, и вообще не были размечены, и стол был размечен как локация, а ящик – нет. Читатели как бы решают, сколько внимания они должны уделить объекту, сравнивают его с другими объектами в рассказе, которые получили аналогичный тег.

Еще меньшую согласованность вызывают двери, проемы, арки и любые другие формы перехода из одного пространства в другое, где задерживаются персонажи. В этот момент переход из одного текстового мира в другой как бы откладывается до некоего решительного действия. При этом в нашем проекте закреплено в правилах разметки, что отдельный текстовый мир создают каналы связи (телефон, видеозвонок) и трансляции (радио, телевидение). Когда персонажи начинают действовать внутри этого канала, то место, в котором они находятся, соединяется с точкой, откуда приходит сигнал, образуя хронотоп, отличный от этих мест, взятых по отдельности.

Несогласованность в примере выше возникает и в разметке словосочетания «простой клерк». Рядом с ним расположено сочетание «чиновники да высокопоставленные особы», в котором довольно легко читается беглое упоминание некоего класса, широкого круга лиц, которому противопоставляет себя герой. В то же время клерк – это его профессия. Идентичны ли категориальные признаки их носителю? Ведь клерка можно рассматривать как упоминание класса людей, профессии вообще. Разметчики решают этот вопрос по-разному, и на данный момент эти расхождения не регулируются в наших правилах разметки.

Несогласие возникает и в отношении животных и объектов неживой природы, которые действуют как персонажи. Место (планета Марс в рассказе Дика, Север в сказке «Дочь Севера») может восприниматься полноценным участником событий. При этом не обязательно возникает анимизм. В примере «Когда машина доставила его домой, в жилой квартал Чикаго…» машину можно рассматривать и как место, и как водителя, который довез героя до Чикаго. Во втором случае «водитель» может даже быть контекстуальным синонимом слова «машина». Спорными становятся случаи, когда не совсем ясно, достаточно ли предмет или место анимированы, одушевлены, стоит ли рассматривать их как отдельных персонажей.

В случае с животными в сказке «Лисичка-сестричка и волк» похожая ситуация возникала при упоминании рыбалки: «поехать за рыбой» – это и двигаться к месту рыбалки, и ловить рыбу, которую можно отметить как группового персонажа или не отмечать, посчитав предметом (в нашем проекте предметы пока не размечаются). Также место может быть частью фразы, указывающей на персонажа: как лексема «Север» в словосочетании «дочь Севера». Разметчики сами решают, как им тегировать такие случаи.

У метрики «альфа Криппендорфа» есть рекомендации относительно того, какой уровень согласия считать достаточным. Для лингвистической разметки в связи с тем, что лингвистическое чутье очень разнится, исследователи допускают чуть более низкие значения, чем обычно. Например, A. Jean-Yves, J. Villaneau, A. Lefeuvre (2014) пишут, что значение от 0.57 до 0.8 можно считать достаточным. В нашем случае самые высокие показатели метрики были зафиксированы для тегов «место» (0.82) и «персонаж» (0.8) в сказке «Лисичка-сестричка и волк». Причем эти показатели даже выше достаточного, а также это те теги, которые разметчики ставили чаще всего. Тег «персонаж» стал самым частым и получил наибольший уровень согласия и в сказке «Дочь Севера» (0.54). Представляется, что персонаж – это наиболее стабильно определяемый элемент текстового мира. Либо он вербализован более явно, чем остальные, либо сами читатели лучше всего координируют свои представления о сюжете вокруг этого персонажа, а не места и времени действия. А может, оба эти фактора влияют на интерпретацию одинаково.

Относительно определения места действия читатели не склонны соглашаться настолько хорошо, чтобы говорить о достаточном уровне. В сказке «Дочь Севера» альфа = 0.28, хотя в «Лисичке…» ее значение – 0.82. Возможно, в некоторых произведениях место играет меньшую роль в повествовании либо подвергается трансформациям (например, становится действующим лицом, как в примерах выше). Также нам стоит проверить, повлияет ли на согласованность ужесточение правил разметки с учетом типичных примеров, которые мы проанализировали выше.

Категория времени оказалась самой неоднозначной. Во всех произведениях она получила низкий уровень согласованности. Мы связываем это с особенностями нарратива, в котором время подразумевается, но не вербализуется так точно, как это происходит с персонажами и локациями. Возможно, время лучше проявлено через последовательность действий, которые совершают персонажи. Это предстоит проверить, когда мы добавим тег для действий в нашу разметку.

Заключение

В результате исследования можно прийти к следующим выводам. Согласованность читателей при разметке таких элементов текстовых миров, как время, место действия и персонажи, редко достигает уровня, который принято считать высоким при оценке надежности разметки. Это связано со спецификой художественного текста, который порождает множество интерпретаций. В таком случае можно рассматривать согласованность как голосование экспертов и выбирать тот тег, который поставило большинство читателей.

Типичные случаи несогласия связаны, например, с ситуациями, когда у читателя мало информации относительно места или персонажа. Также некоторые объекты живой и неживой природы могут быть восприняты как действующие лица либо вместилища в зависимости от важности их роли в сюжете. И, наконец, слова, обозначающие категории лиц, например профессии, могут восприниматься как указатели на персонажа, который относится к этой категории.

Согласованность читателей варьируется в зависимости от текста, из-за чего мы можем предположить, что в некоторых текстах элементы текстовых миров выражены более однозначно. Самое высокое согласие читателей было обнаружено для категорий места и персонажей. Категория времени получила самые низкие оценки согласованности и в целом представлена самым малым количеством тегов в разметке. Видимо, художественное время менее вербализовано и больше привязано к действиям, событиям сюжета.

В перспективе планируются улучшения процесса разметки, которые способны повысить качество автоматического выявления элементов текстовых миров в художественных текстах. На данный момент проводится работа над расширением разметки и правилами аннотации. Типичные случаи, где мнения читателей могут разойтись, планируется отмечать особым тегом, чтобы отделить, какие группы элементов выражены более явно и, следовательно, лучше подходят для автоматического выявления. Также мы добавим в разметку объекты неживой природы и категорию действия, вербализованную через глаголы и глагольные группы, что, возможно, заменит или дополнит категорию времени. Планируется отделить случаи, когда в разметке может быть однозначная интерпретация (например, если топоним входит в имя персонажа, то размечать это только как имя; см. пример с «дочь Севера»), от случаев вариации, которая отражает субъективный опыт чтения. В результате мы ожидаем получить более подробное описание художественного нарратива и, возможно, успешную модель автоматического выявления его элементов.

Источники | References

1. Бахтин М. М. Вопросы литературы и эстетики. М.: Худ. лит., 1975.
2. Евсеев О. В., Кох А. Н., Михалькова Е. В. Сопоставительный анализ элементов текстовых миров в литературной сказке «Золушка» Ш. Перро (в переводе на русский) и одноименном киносценарии Е. Шварца // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. 2023. Т. 9. № 1 (33).
3. Олейник А. Н., Попова И. П., Кирдина С. Г., Шаталова Т. Ю. Надежность и достоверность в контент-анализе текстов: выбор показателей // Психологический журнал. 2014. Т. 35. № 6.

4. Bakeman R., Quera V., McArthur D., Robinson B. F. Detecting sequential patterns and determining their reliability with fallible observers // Psychological Methods. 1997. Vol. 2 (4).
5. Beck C., Booth H., El-Assady M., Butt M. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias // Proceedings of the 14th Linguistic Annotation Workshop. Barcelona, 2020.
6. Bell A., Ryan M. L. Possible Worlds Theory and Contemporary Narratology. Lincoln: University of Nebraska Press, 2019.
7. Bird S., Klein E., Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Sebastopol: O'Reilly Media, Inc., 2009.
8. Cohen J. A coefficient of agreement for nominal scales // Educational and Psychological Measurement. 1960. Vol. 20 (1).
9. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit // Psychological Bulletin. 1968. Vol. 70 (4).
10. Detkova J., Novitskiy V., Petrova M., Selegey V. Differential semantic sketches for Russian internet-corpora // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". 2020. Vol. 19.
11. Fleiss J. L. Measuring nominal scale agreement among many raters // Psychological Bulletin. 1971. Vol. 76 (5).
12. Gavins J. Text World Theory: An Introduction. Edinburgh: Edinburgh University Press, 2007.
13. Gibbons A., Whiteley S. Do worlds have (fourth) walls?: A Text World Theory approach to direct address in Fleabag // Language and Literature. 2020. Vol. 30 (2).
14. Gwet K. L. Chapter 6 // Gwet K. L. Handbook of Inter-Rater Reliability. Gaithersburg: Advanced Analytics, LLC, 2014.
15. Hayes A. F., Krippendorff K. Answering the call for a standard reliability measure for coding data // Communication Methods and Measures. 2007. Vol. 1 (1).
16. Ho Y., Lugea J., McIntyre D., Wang J., Xu Z. Projecting (un)certainty: A text-world analysis of three statements from the Meredith Kercher murder case // English Text Construction. 2018. Vol. 11 (2).
17. Ho Y.-F., Lugea J., McIntyre D., Xu Z., Wang J. Text-world annotation and visualization for crime narrative reconstruction // Digital Scholarship in the Humanities. 2019. Vol. 34 (2).
18. Honnibal M., Montani I., Van Landeghem S., Boyd A. spaCy: Industrial-Strength Natural Language Processing in Python // Zenodo. 2020. https://dx.doi.org/10.5281/zenodo.1212303
19. Hu Y., Mao H., McKenzie G. A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements // International Journal of Geographical Information Science. 2019. Vol. 33 (4).
20. Jean-Yves A., Villaneau J., Lefeuvre A. Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: Experimental studies on emotion, opinion and coreference annotation // Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, 2014.
21. Krippendorff K. Content Analysis: An Introduction to Its Methodology. Los Angeles, 2018.
22. Landis J. R., Koch G. G. The measurement of observer agreement for categorical data // Biometrics. 1977. Vol. 33 (1).
23. Mikhalkova E., Protasov T., Drozdova A., Bashmakova A., Gavin P. Towards annotation of text worlds in a literary work // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". 2019. Vol. 18.
24. Mikhalkova E., Protasov T., Gavin P., Bashmakova A., Drozdova A. Modelling narrative elements in a short story: A study on annotation schemes and guidelines // Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, 2020.
25. Peng S., Sun Z., Loftus S., Plank B. Different tastes of entities: Investigating human label variation in named entity annotations // The Third Workshop on Understanding Implicit and Underspecified Language. Malta, 2024.
26. Raghunath R. Possible Worlds Theory and Counterfactual Historical Fiction. Cham: Springer Nature, 2020.
27. Sang Y., Mou X., Li J., Stanton J., Yu M. A survey of machine narrative reading comprehension assessments // 31st International Joint Conference on Artificial Intelligence. Vienna: IJCAI, 2022.
28. Sirinarang B., Wijitsopon R. A cognitive stylistic approach to mind style in the memoir man's search for meaning // Journal of Studies in the English Language. 2021. Vol. 16 (1).
29. Srivatsa S., Srinivasa S. Narrative plot comparison based on a bag-of-actors document model // Proceedings of the 29th ACM Conference on Hypertext and Social Media (HT'18) / Association for Computing Machinery. N. Y., 2018.
30. Stockwell P. Cognitive Poetics: An Introduction. Abingdon-on-Thames: Routledge, 2020.
31. Tinsley H. E., Weiss D. J.Interrater reliability and agreement of subjective judgments // Journal of Counseling Psychology. 1975. Vol. 22 (4).
32. Uma A., Fornaciari T., Dumitrache A., Miller T., Chamberlain J., Plank B., Simpson E., Poesio M. SemEval-2021 Task 12: Learning with Disagreements // Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). 2021. https://doi.org/10.18653/v1/2021.semeval-1.41
33. Wang J., Ho Y.-F., Xu Z., McIntyre D., Lugea J. The visualisation of cognitive structures in forensic statements // 20th International Conference Information Visualisation (IV). Lisbon: IEEE, 2006. https://doi.org/10.1109/IV.2016.60
34. Weber-Genzel L., Peng S., De Marneffe M. C., Plank B. VariErr NLI: Separating annotation error from human label variation // arXiv. 2024. https://doi.org/10.48550/arXiv.2403.01931
35. Werth P. Accommodation and the myth of presupposition: The view from discourse // Lingua. 1993. Vol. 89 (1).
36. Werth P. Extended metaphor – a text-world account // Language and Literature. 1994. Vol. 3 (2).
37. Werth P. Text Worlds: Representing Conceptual Space in Discourse. Harlow: Longman, 1999.

Информация об авторах | Author information

RU

Михалькова Елена Владимировна[1], к. филол. н.
[1] Европейский университет в Санкт-Петербурге

EN

Elena Vladimirovna Mikhalkova[1], PhD
[1] European University at Saint Petersburg

[1] *e.mikhalkova@eu.spb.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 03.08.2024; опубликовано online (published online): 16.09.2024.

Ключевые слова (keywords): нарративные категории; теория текстовых миров; согласованность читателей; разметка художественного текста; метрика согласованности; надежность разметки; narrative categories; Text World Theory; inter-rater agreement; annotation of a literary text; agreement metric; annotation reliability.