ru:corner

Разметка именованных сущностей и кореференции в brat

Анастасия Никифорова
steysie@gmail.com

Сергей Терновых
fostroll@gmail.com

Денис Киреев
dkireev.71@gmail.com

Константин Ремизов
mr.enslin@mail.ru

Январь 2021

Содержание

1 Окно разметки в brat **1**

1.1 Начало работы. Авторизация 1

1.2 Выбор текста из коллекции 2

1.3 Выделение и разметка текста 2

1.4 Связи между размеченными упоминаниями 3

1.5 Изменение и удаление меток 4

1.6 Нерелевантные тексты 4

1.7 Поиск незнакомых терминов в Google и Википедии 5

1.8 Комментарии к сущностям 5

2 Разметка именованных сущностей в brat **5**

2.1 Алиасы названий сущностей в интерфейсе 6

2.2 Вложенные и пересекающиеся сущности 6

2.3 Атрибуты типов сущностей 10

2.4 Различия подтипов сущностей Time и Date 11

3 Разметка кореференции в brat **11**

4 Отдельные случаи **13**

4.1 Уточняющие прилагательные 14

4.2 Предлоги 14

4.3 Знаки препинания 14

4.4 Кавычки 15

4.5 Временные выражения с неопределенными границами 16

4.6 Одиночные именованные сущности и анафора 16

4.7 Numeric: Quantity vs. Cardinal в случае эллипсиса 17

4.8 Выделение сущности, разбитой на две строки 17

4.9 Возможные ошибки и предупреждения 17

Приложение A Changelog **19**

1 Окно разметки в brat

brat – инструмент для разметки именованных сущностей (и других текстовых интервалов) и связей между ними. В ru:corner brat используется для разметки именованных сущностей и кореференции.

Окно разметки в brat изображено на Рисунке 1.

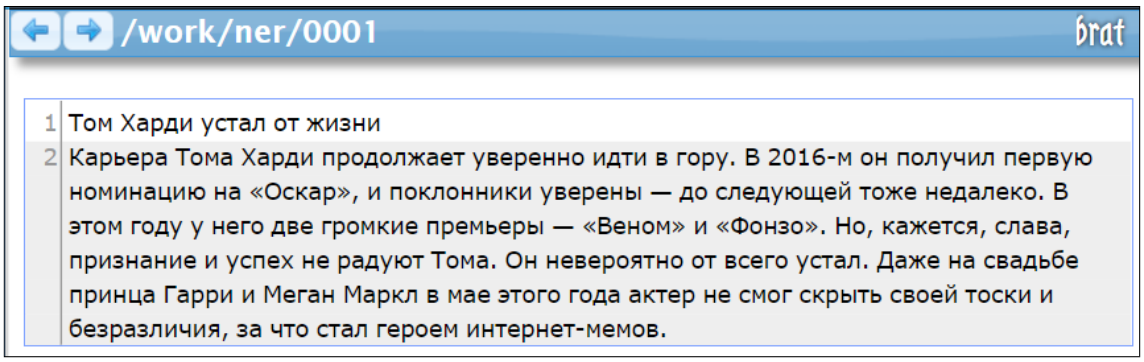

Рис. 1: Пример окна разметки в brat

1.1 Начало работы. Авторизация

Разметка в brat доступна только для авторизованных пользователей. Чтобы авторизоваться, наведите курсор к шапке страницы, над текстом. В появшейся строке выберите **Login** (Рисунок 2).

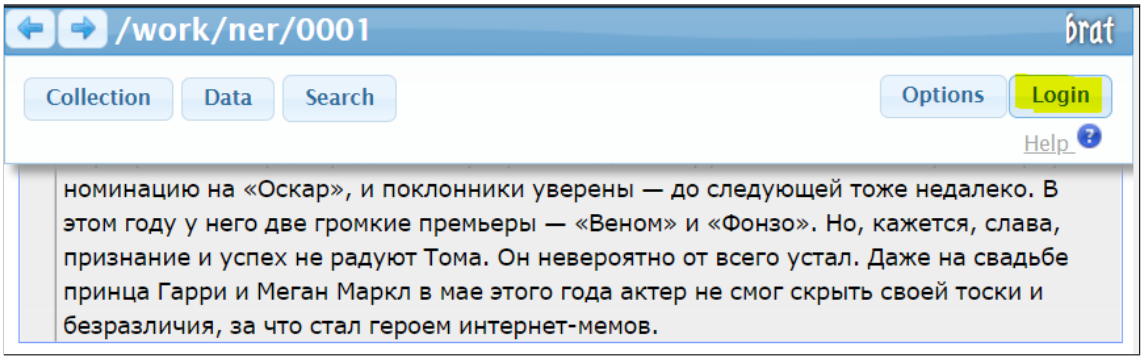

Рис. 2: Авторизация в brat

В появившемся окне введите логин и пароль и нажмите **OK** (Рисунок 3). После успешной авторизации внизу страницы появится приветственное сообщение.

После входа в аккаунт можно выбрать текст из коллекции и приступать к разметке именованных сущностей или кореференции.

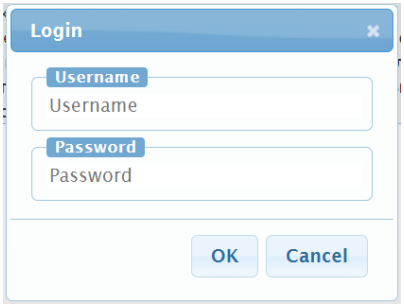

Рис. 3: Окно авторизации

1.2 Выбор текста из коллекции

Документы для разметки являются частью коллекции (Collections). Для выбора первого текста разметки, пройдите в нужную директорию и выберите файл двойным нажатием (Рисунок 4). Откроется окно с текстом как на Рисунке 1.

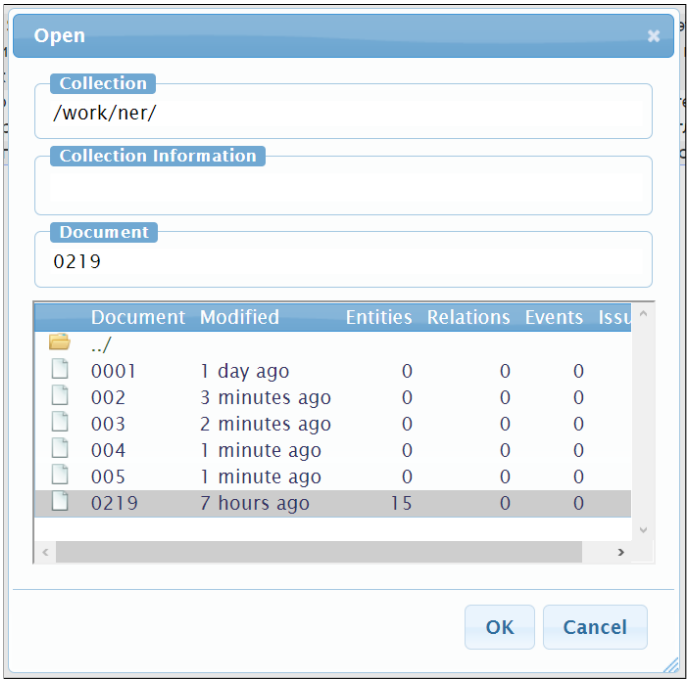

Рис. 4: Пример окна Collections

1.3 Выделение и разметка текста

Для того, чтобы разметить отрезок текста как именную сущность или упоминание, выделите курсором нужное слово или фразу от первой буквы первого слова, до последней буквы последнего слова. Единичные слова можно выделять, нажав на них дважды левой кнопкой мыши.

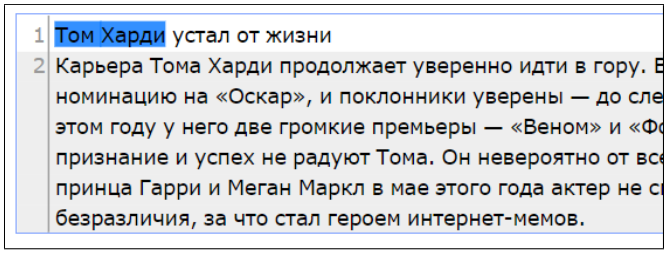

Рис. 5: Пример выделенного отрезка текста

Как только текст выделен, появится окно выбора типа сущности (Рисунок 6). Подробнее о видах меток отдельно для разметки именованных сущностей и кореференции можно прочитать в разделах 2 и 3.

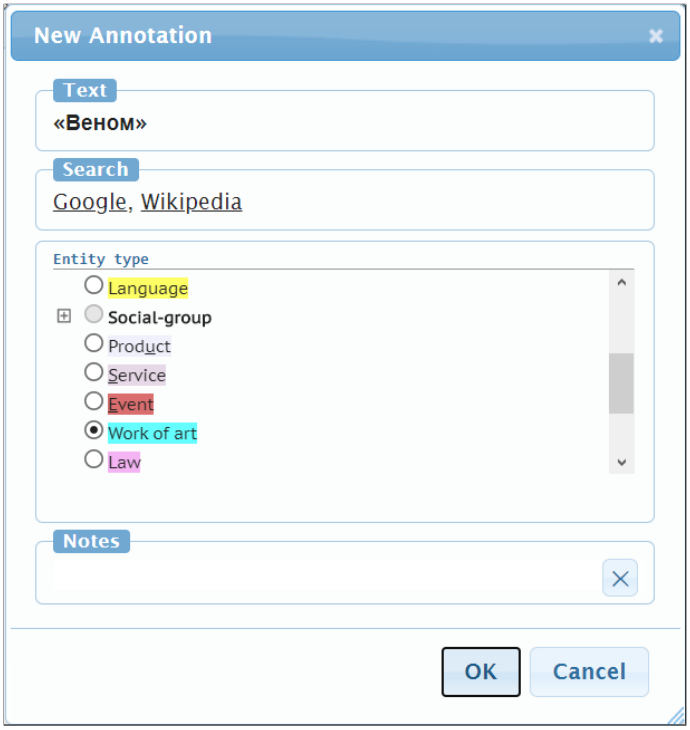

Рис. 6: Пример окна выбора метки именованной сущности

В окне выбора типов сущностей выберите подходящую метку и нажмите `Enter` или щелкните курсором по **OK**.

Когда весь текст размечен, нажмите клавишу → или кликните на кнопку ⇒ в левом верхнем углу `brat`.

1.4 Связи между размеченными упоминаниями

При разметке кореференции необходимо попарно связать размеченные упоминания стрелкой и выбрать тип связи.

Для образования связи, нажмите на метку одного из упоминаний (часто – зависимый член), как на Рисунке 7 и перетащите появившуюся стрелку на второе упоминание, как на Рисунке 8.

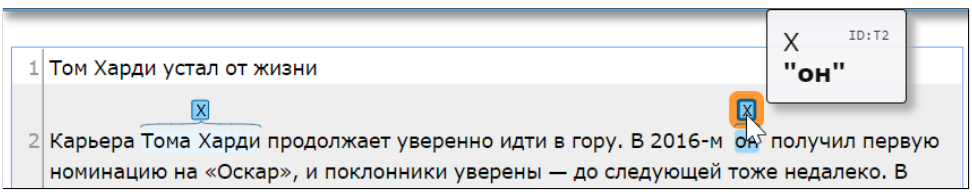

Рис. 7: Шаг 1. Выбор метки одного из упоминаний

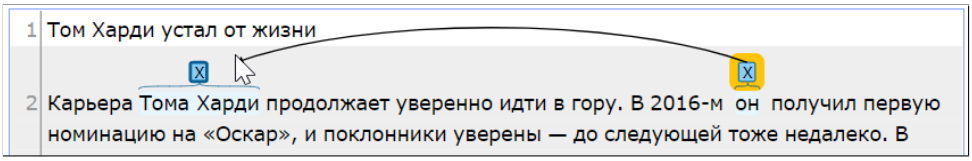

Рис. 8: Шаг 2. Связывание кореферентной группы

Когда упоминания связаны, появится окно выбора типа связи. Подробнее о видах кореферентных связей можно прочитать в разделе 3.

После выбора тип связи отобразится на стрелке (Рисунок 9).

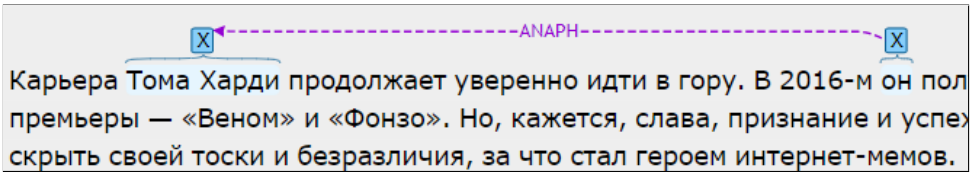

Рис. 9: Шаг 3. Выбор связи

1.5 Изменение и удаление меток

Чтобы изменить или удалить метку сущностей (в случае ошибочного выбора метки и т. п.), дважды щелкните на название метки. Появится окно выбора типа сущностей (Рисунок 10).

Для изменения типа сущности, выберите другую метку из списка.

Для изменения выделенного диапазона, нажмите **Move** и заново выделите сущность.

Для удаления метки, нажмите **Delete**.

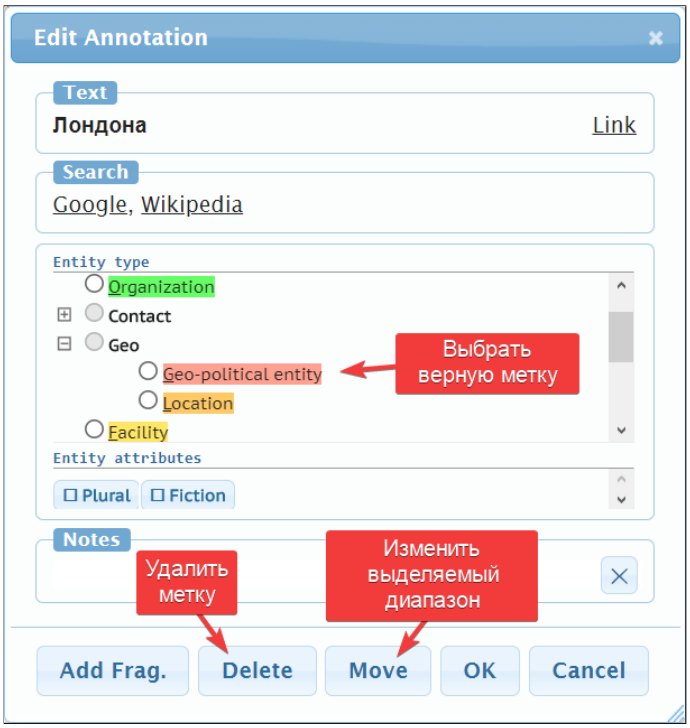

Рис. 10: Изменение и удаление меток именованных сущностей

1.6 Нерелевантные тексты

В коллекции могут попадаться тексты, в которых преобладают другие языки, которые также используют кириллицу – например, украинский или белорусский. Также могут встречаться тексты, преимущественно состоящие из стихов.

В таких текстах необходимо отметить только первое слово специальным тегом `!!! INVALID DOCUMENT !!!`. Весь остальной текст следует оставить неразмеченным, как показано на Рисунке 11.

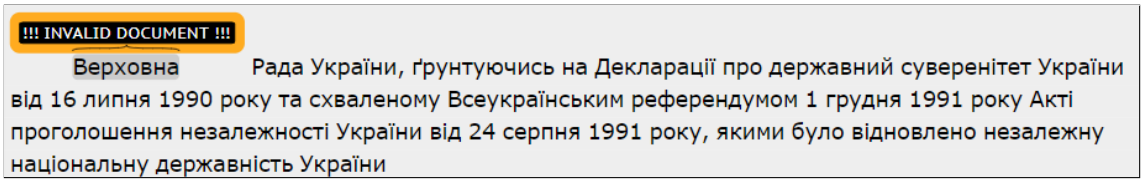

Рис. 11: Пример разметки нерелевантного текста

Если в тексте преобладает русский язык, но встречаются фразы на других языках, такой текст размечается как обычно.

1.7 Поиск незнакомых терминов в Google и Википедии

Чтобы найти значение незнакомых слов в Google и в Википедии, выделите нужный отрезок текста. В появившемся окне в разделе **Search** нажмите на **Google** или **Wikipedia** (Рисунок 12).

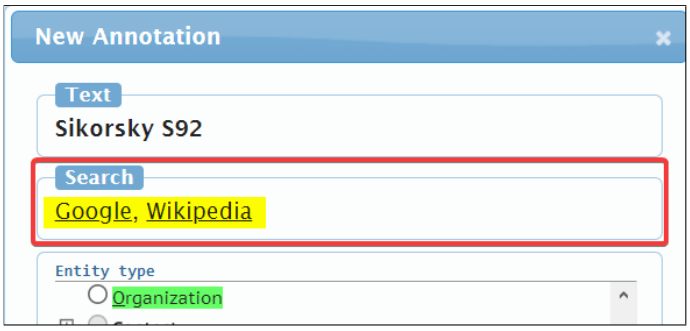

Рис. 12: Поиск незнакомых терминов в Google и Wikipedia

1.8 Комментарии к сущностям

Если вы не уверены, правильно ли выделена сущность, можно оставить короткий комментарий в окне выбора метки в разделе **Notes**. Чтобы удалить комментарий, нажмите на крестик справа в поле **Notes**.

2 Разметка именованных сущностей в `brat`

Для разметки именованных сущностей в ru:corner предусмотрено 44 различных типа сущностей. Подробное описание сущностей с примерами приведено ниже в Таблице 1.

Некоторые сущности могут иметь дополнительные атрибуты, один или несколько из следующих категорий: *Plural*, *Fiction*, *Unconsciuos*, *Department*, *Media*, *Citizenship*, *Project*, *Trademark*. Подробнее об атрибутах – в Разделе 2.3.

Для наиболее частотных типов сущностей предусмотрены горячие клавиши (колонка Hotkey в Таблице 1). При открытом окне выбора типа именованных сущностей нажмите подходящую горячую клавишу и нажмите `Enter`. Над выделенным упоминанием появится нужный тип сущности.

2.1 Алиасы названий сущностей в интерфейсе

В зависимости от длины выделенной фразы, в интерфейсе могут использоваться укороченные варианты одного и того же типа сущности, как показано на Рисунке 13.

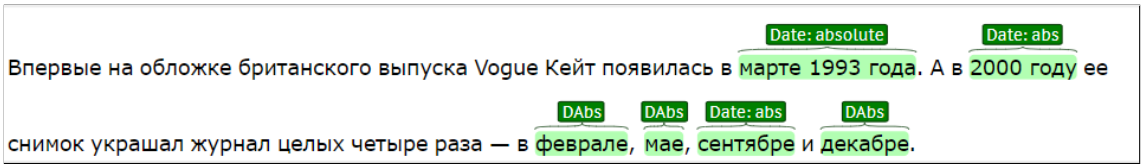

Рис. 13: Примеры алиасов сущности `Date: absolute`

2.2 Вложенные и пересекающиеся сущности

Именованные сущности могут быть составными, то есть могут включать в себя вложенные сущности (а-ля “матрёшка”). Примеры разметки с учетом вложенности на Рисунке 14.

Примечание. Пересекающимися могут быть только сущности из категории `Quantity` (количество). Например, “{шестерых [членов} экипажа]”

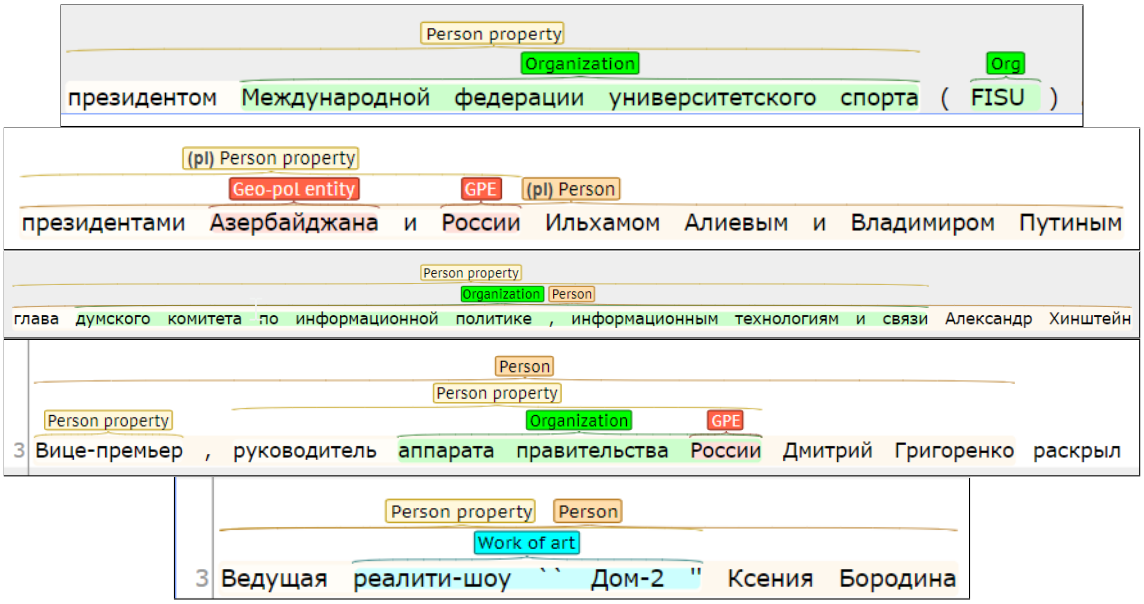

Рис. 14: Примеры пересекающихся и вложенных сущностей

Любые типы сущностей могут быть составными. Чаще других составными оказываются сущности *Person*, которые часто включают в себя сущности *Person Property*, *GPE* и т. д.

Таблица 1. Типы сущностей, их описания и примеры

Категория	Тип сущности	Атрибуты	Hotkey	Описание типа сущности	Пример
!!! INVALID DOCUMENT !!!				Любые стихи и тексты, написанные преимущественно не на русском языке (украинский, белорусский и др.)	*Отметьте **первое слово** в нерелевантном тексте*
Person		☐ Plural ☐ Fiction ☐ Unconscious **[default]**	P	Имя или псевдоним реального человека. *В качестве Person могут употребляться не только имена собственные, но и другие устойчивые выражения, зачастую имя+свойство, если это выражение часто используется в СМИ, кино или литературе для идентификации этого известного человека или персонажа.*	*Смирнов Иван Петрович, Сан Саныч, Машуня, Андрюха, Петр I, Бейонсе, Роберт Дауни Младший, Шакил О'Нил, Шекспир, Пушкин, Моцарт, 50 cent, Моргенштерн, королева Елизавета, Президент Путин, товарищ Сталин, дорогой Леонид Ильич, президент Борис Ельцин*
		☑ Plural		Имена нескольких людей в одной сущности	*Павел и Марина Смирновы, Дэвид и Виктория Бэкхам*
		☑ Fiction		Имена или псевдонимы вымышленных персонажей, богов (кино, литература, комиксы, религия, мифология и т. п.)	*Евгений Онегин, Губка Боб, попугай Кеша, Энакин Скайуокер, Иисус Христос, Будда, Мальчик который выжил, Тот-чье-имя-нельзя-называть*
		☑ Unconsciuos		Любые имена, которыми люди называют неразумные объекты: животные, техника и т. п.	*Барсик, Белка, Стрелка, Несси (лох-несское чудовище), Флиппер (дельфин), "Сметливый" (корабль)*
Person property		☐ Plural ☐ Fiction **[default]**	R	Чин, звание, титул, должность, профессия и т. п.	*генерал-майор, королева, канцлер, сварщик, менеджер, президент, инженеры, разработчики, депутаты, ученые, исследователи, срочник, заключенные, задержанные*
		☑ Plural		Несколько различных Property в рамках одной сущности	*разработчики и дизайнеры проекта, несколько строителей и электриков*
		☑ Fiction		Вымышленные титулы/профессии и т. п.	*профессор защиты от темных искусств, штурмовик Первого Ордена*
Organization		☐ Plural ☐ Fiction ☐ Department **[default]**	O	Компании, агентства, издательства, радиостанции, институты, политические партии и т. п., имеющие орг. структуру	*Ростелеком, Amazon, ПАО "Газпром", партия "Единая Россия", РПЦ, компания Орифлейм, издательство ЭКСМО, Эхо Москвы, Белорусская оппозиция*
		☑ Plural		Несколько организаций или отделов в рамках одной сущности	*Норильский и Алтайский горнодобывающие заводы, финансовый и юридический отделы, Московская и Екатеринбургская епархии*
		☑ Fiction		Вымышленные организации (кино, литература и т. п.)	*галактический сенат, Stark Industries, Спектр, школа волшебства Хогвартс*
		☑ Department		Отделы внутри организаций	*IT-департамент, отдел кадров, совет директоров, топ-менеджмент*
Contact	**Address**		A	Адрес местоположения, зачастую - город, улица, дом, квартира, индекс	*123456, Москва, Тверская улица, дом 10, строение 2; адрес: ул. Ленина, д. 5*
	Phone		H	Номер телефона	*+7 (123) 456-78-90, тел. 81234567890, телефон 44-22-33*
	Email		M	Адрес электронной почты	*name@wsite.com, email: user@edu.site.org*
	Web address			Ссылка на веб-страницу	*https://website.com/info/, google.com*
	Other-contact			Имя пользователя, профиль в instagram и т. п.	*@someusername, telegram: @bestname, tg username, instagram: followme*
Event		☐ Plural **[default]**	E	Ураганы, сражения, войны, спортивные состязания, праздники и т. п. (не одиночные слова типа "концерт", "церемония" и т. д.)	*Новый год, ураган Катрина, Чемпионат мира по футболу, Брусиловский прорыв, предизентские выборы в США, концерт Ольги Бузовой, концерт Баскова и Киркорова (одно мероприятие)*
		☑ Plural		Несколько Event в рамках одной сущности	*ураганы Катрина и Рита, концерты Баскова и Киркорова (разные мероприятия)*
Geo	**Geo-political entity (GPE)**	☐ Plural ☐ Fiction **[default]**	G	Географическая зона, имеющая политическую структуру + космические станции (GPE)	*Ростов-на-Дону, Швейцария, Ближний Восток, г. Москва, СНГ, СССР, Катманду, деревня Кунцево, Алтуфьевский район, столица (если понятно, что Москва)*
		☑ Plural		Несколько GPE	*Московская и Ленинградская области*
		☑ Fiction		Вымышленная географическая зона	*Атлантида, Хогвартс, Нарния, Вестерос, Средиземье, Готэм-сити, Асгард*
	Location	☐ Plural ☐ Fiction **[default]**	L	Места, природные: горные цепи, водоемы + планеты, галактики, созвездия, кометы и т. п.	*озеро Байкал, Волга, р. Нева, оз. Чад, Гималаи, Эверест, Земля, Луна, Марс, Астероид 501647, Млечный путь*
		☑ Plural		Несколько Location	*реки Тигр и Евфрат*
		☑ Fiction		Вымышленные места	*Мглистые горы, Андуин, река Яруга*

Таблица 1. Типы сущностей, их описания и примеры

Категория	Тип сущности	Атрибуты	Hotkey	Описание типа сущности	Пример
Facility		☐ Plural ☐ Fiction **[default]**	F	Достопримечательности, здания, аэропорты, шоссе, мосты, улицы, площади, переулки и т. п.	*памятник Ильичу, Внуково, Трасса М4, ТРЦ "Авиапарк", Дворцовый мост, Арбат, Трехсвятская улица, проспект Ленина*
		☑ **Plural**		Несколько Facility	*трассы М4 и М5*
		☑ **Fiction**		Вымышленные Facility (кино, литература и т. п.)	*Замок Саурона, Замок Дарта Вейдера, Галактический рынок*
Language		☐ Plural ☐ Fiction **[default]**		Языки естественные или искусственные (не из кино/литературы)	*русский язык, итальянский, иврит, хинди, язык йоруба, урду, эсперанто, северные диалекты русского языка*
		☑ **Plural**		Несколько объектов Language в рамках одной сущности	*южные и восточные диалекты, тюркские и финноугорские языки, говоры и наречия русского языка*
		☑ **Fiction**		Вымышленные языки (кино, литература)	*клингон, на'ви, дотракийский, новояз*
Social-group	**Nationality**	☐ Citizenship ☐ Plural **[default]**		Принадлежность к нации по происхождению, рождению или иным образом	*русские, итальянец, гречанка, папуасы, финны, афроамериканцы, кореец, индус, армянского происхождения*
		☑ **Citizenship**		Гражданство какой-либо страны	*граждане РФ, гражданка Эстонии, жители Китая, американцы*
		☑ **Plural**		Несколько разных объектов Nationality в одной сущности	*Граждане Грузии и Армении, жители Северной и Южной Кореи*
	Family			Обозначения родственных связей	*мать, брат, старший сын, сводная сестра, двоюродная тетя, брат жены*
	Religious group	☐ Plural **[default]**		Принадлежность к определенной религии	*православные, католик, старообрядцы, амиши, шииты, буддист, пастафарианцы*
		☑ **Plural**		Несколько разных религиозных групп в одной сущности	*православные, католические и протестантские христиане*
	Political group	☐ Plural **[default]**		Принадлежность к политической группе	*республиканцы, члены партии "Единая Россия", национал-демократы, коммунисты, члены партии «Свобода»*
		☑ **Plural**		Несколько политических групп в одной сущности	*члены КПРФ и ЛДПР, члены партий Единая Россия и Яблоко*
	Other group	☐ Plural ☐ Fiction **[default]**		Другие социальные группы, члены которых имеют что-то общее	*безработные, руководители, веганы, гангстеры, новые русские, готы, геи, бомжи, преступник, избиратели, бюджетники, поклонники, эксперты*
		☑ **Plural**		Несколько разных групп в одной сущности	*представители готов и эмо, участники клубов рукоделия и гончарства*
		☑ **Fiction**		Вымышленные социальные группы	*хоббиты, эльфы, орки, джедаи, члены банды "Железные рукава"*
Product		☐ Plural ☐ Project ☐ Trademark **[default]**	U	Электроника, автомобили, оружие, продукты питания, одежда и т. п. (конкретные продукты с названием/производителем)	*пылесос Dyson, диваны Икеа, Range Rover, Айфон, Орион чокопай, пистолет Макарова, АК-47, косметика Орифлейм, платье бренда Виктория Бекхам*
		☑ **Plural**		Несколько продуктовых сущностей в рамках одной	*диваны Икеа и Hoff, пылесосы Dyson и Xiaomi, косметика Эйвон и Орифлейм*
		☑ **Project**		Проекты, программы (государственные, научные, космические и т. п.)	*мегасаенс-проект NICA, нацпроект "Наука", программа обмена студентами Erasmus +, программа «Вояджер»*
		☑ **Trademark**		Бренд, торговая марка продукта (не в значении "компания/организация")	*Dyson, Apple, Орифлейм, Виктория Бекхам, Шанель, Tesla*
Service		☐ Plural **[default]**	S	Различные предприятия, приложения и технологии предоставления услуг	*чистка обуви, ремонт ноутбуков, доставка еды, телефония, Яндекс.Еда*
		☑ **Plural**		Несколько сервисов в одной сущности	*услуги прачечной и химчистки, услуги телефонии и интернета*
Work of art				Названия книг, песен, картин, фильмов, ТВ шоу, сериалов и т. п.	*Властелин колец, Мона Лиза, "Белые розы", Черный квадрат Малевича*
Law		☐ Plural **[default]**		Нормативно-правовые акты	*Конституция, статья 20.6.1 КоАП, Закон о защите прав потребителей*
		☑ **Plural**		Несколько Law в одной сущности	*п. 2 статьи 228 и пп. 3 и 4 статьи 230 УК РФ*

Таблица 1. Типы сущностей, их описания и примеры

Категория	Тип сущности	Атрибуты	Hotkey	Описание типа сущности	Пример
Date	**Date: absolute**		D	Явная дата (когда? какой день/месяц/год?)	*20 октября 2000 года, пятое августа, январь 2021 года, 2020, в сентябре*
	Date: relative			Относительная дата, включая конечные предлоги (когда относительно другой даты?)	*через сутки, три дня назад, прошлый год, по прошествии двух лет, на следующий день после*
	Date: period			Период, измеряемый в днях, неделях, месяцах и т. д.	*с 5 января по 10 февраля, Мезозойская Эра, Эпоха Возрождения, за сентябрь, за последние шесть лет*
	Date: duration			Продолжительность, измеряемая в днях, неделях, годах и т. д. (как долго?)	*в течение двух недель, три года, пять месяцев и два дня, за сутки, полдня*
Time	**Time: absolute**		T	Явное время (когда? во сколько?)	*15:30, полтретьего, пятнадцать минут одиннадцатого, три часа*
	Time: relative			Относительное время, включая конечные предлоги (когда относительно другого времени?)	*два часа назад, через минуту, за два часа до, по прошествии двух часов, в течение трех часов (напр., будет закончена работа, выполнен заказ и т.п.)*
	Time: period			Временной период, измеряемый в секундах, минутах, часах	*с двух до пяти, с 12-ти до 23-х, за последние два с половиной часа*
	Time: duration			Продолжительность времени, измеряемая в секундах, минутах, часах (как долго по времени?)	*в течение трех часов, пару минут, три часа, за полчаса*
Numeric	**Ordinal**		3	Числа (и слова) для подсчета предметов (какой по счету?)	*первый, второй, тысячный, следующий, предыдущий, (пред)последний*
	Money		4	Денежная сумма, включая название валюты (сколько?)	*$500, два евро, 100 рублей, один тенге, 300 динаров, 55 франков, сорок пять рублей двадцать две копейки*
	Percent		5	Процент (включая “%”)	*десять процентов, полпроцента, 99.9%, 147%, на 30.1% меньше*
	Age		6	Возраст человека или предмета	*25 лет, 20-летний, годовалый, новорожденный*
	Quantity		8	Измерения, количество чего-либо с единицей измерения (сколько?)	*два кило, два стакана, 3 ч. л., 670 км, полтора метра, 6.4 Вт, метровый, более трех человек, около пяти видов, ни разу, один из, дважды, вчетверо реже, однажды (в значении "однократно, один раз")*
	Cardinal		0	Числа, не относящиеся к категориям выше	*один, два, десять, миллион, 532, 12.331, 52 828, две трети, 1/3*
Other term		☐ Plural ☐ Fiction **[default]**		Термины, не подпадающие под категории выше, включая научные термины, болезни, технологии и т. д.	*4G, гигабайт, рибонуклеотид, минорные актиноиды, премия "Оскар", COVID-19, коронавирус, инсульт, ишемия миокарда*
		☑ **Plural**		Несколько разных терминов, объединенных в одну сущность	*гепатит A и B, азотная и соляная кислоты*
		☑ **Fiction**		Вымышленные термины	*Татуин, Пандора, Скайнет, криптонит*

2.3 Атрибуты типов сущностей

Некоторые сущности могут иметь один или несколько атрибутов из следующих:

- **Plural** – два или более разных объекта, объединенных в одну именованную сущность.
- **Fiction** – вымышленная сущность, как правило, относящаяся к литературе, кино, сериалам и др.
- **Unconscious** – имена, которыми люди называют неразумные объекты. Доступно только для типа сущности `Person`.
- **Department** – отдел внутри организации. Доступно только для типа сущности `Organization`.
- **Media** – любые медиа-источники: издательства, социальные сети, телевизионные каналы и т. п. Доступно только для типа сущности `Organization`.
- **Citizenship** – гражданство какой-либо страны. Доступно только для типа сущности `Nationality`.
- **Project** – проекты, программы (государственные, научные, космические и т. п.). Доступно только для типа сущности `Product`.
- **Trademark** – бренд, торговая марка продукта (не в значении *Organization*). Доступно только для типа сущности `Product`.

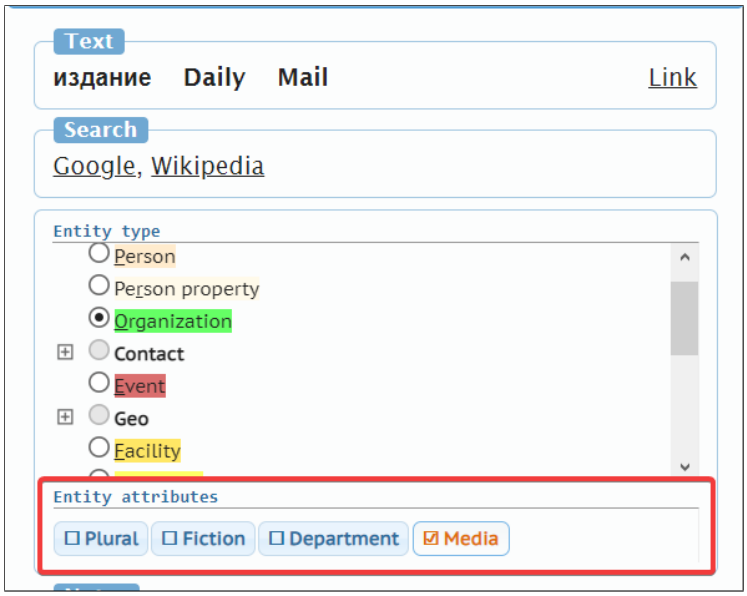

Рис. 15: Выбор атрибутов

По умолчанию ни один атрибут не выбран. Чтобы выбрать атрибут, после выбора типа сущности нажмите на один из атрибутов в разделе **Entity attributes**. В соответствующем поле появится флажок, и атрибут подсветится оранжевым (Рисунок 15).

2.4 Различия подтипов сущностей Time и Date

Time – сущность, измеряемая в часах, минутах, секундах и т. д.

Date – сущность, измеряемая в днях, неделях, месяцах, годах и т. д.

Сущности **Date** и **Time** – только конкретные дата и время. Слова "*недавно, скоро, в последнее время*" и т. п. не выделяются как сущность.

На Рисунке 16 наглядно изображены различия подтипов временных сущностей.

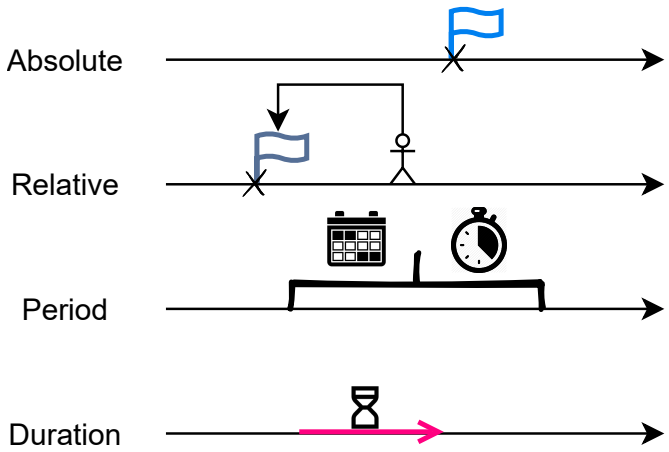

Рис. 16: Различия подтипов Time и Date

Absolute и *Relative* - всегда конкретное, точечное время. *Period* - период времени, который можно ограничить началом и концом. *Duration* - измеряемая продолжительность времени.

3 Разметка кореференции в `brat`

Все упоминания в `brat` отмечаются меткой [X]. На Рисунке 17 показаны выделенные упоминания в тексте.

Все выделенные упоминания нужно попарно связать кореферентными связями. Всего в ru:corner четыре вида кореферентных связей: ANAPH, CATAPH, IDENT и APPOS. Подробную информацию об этих связях можно найти в Руководстве по разметке кореференции в ru:corner.

На Рисунке 18 показан пример текста с размеченными связями между упоминаниями.

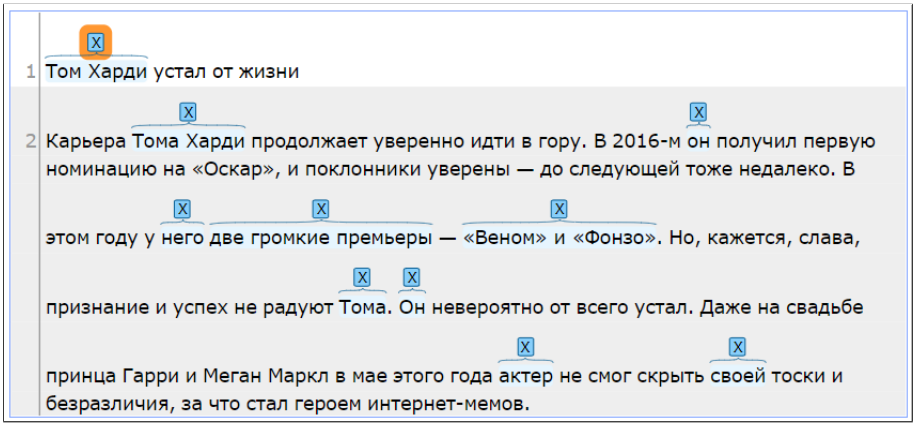

Рис. 17: Пример выделения упоминаний в `brat`

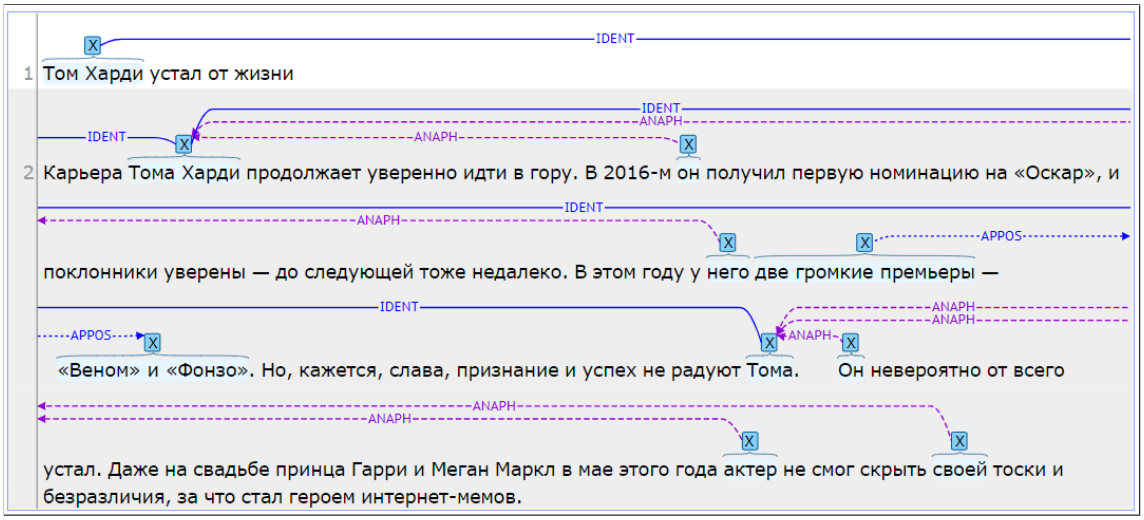

Рис. 18: Пример кореферентной разметки в `brat`

Если в тексте нет референта, и все упоминания выражены местоимениями или нереферентными именными группами, следует отметить все упоминания такой кореферентной цепочки атрибутом **NoRef** (Рисунок 19).

Если у кореферентной группы указан референт в тексте, атрибут **NoRef** выбирать не нужно.

Связь IDENT ненаправленная, связь ANAPH всегда направлена влево, связь CATAPH всегда направлена вправо, а связь APPOS может быть направлена и влево, и вправо.

Для выбора другой метки связи, выберите её из списка.

Примечание. Связь IDENT ненаправленная, ее нельзя переименовать в одну из направленных связей.

Чтобы поменять направление связи между упоминаниями, дважды кликните на нужную стрелку и в появившемся окне выбора связи нажмите **Reverse**.

Чтобы изменить упоминание-цель (упоминание, на которое указывает стрелка), дважды кликните на стрелку, нажмите **Reselect** и переместите освободившуюся стрелку на новое слово.

Рис. 19: Выбор атрибута **NoRef** для нереферентных упоминаний

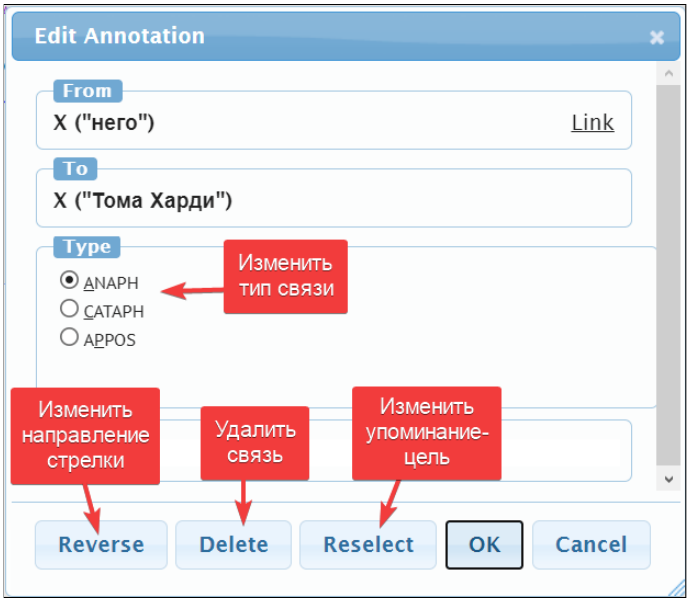

Рис. 20: Изменение или удаление связи между упоминаниями

Для удаления связи, дважды кликните на стрелку и нажмите **Delete** в окне выбора типа связей.

4 Отдельные случаи

В интерфейсе `brat` между всеми токенами[1] для удобства разметки – три пробела. Каждое предложение отображается на новой строке, а между каждым абзацем есть пустая строка. Пример текста изображен на Рисунке 21.

[1]Токены – отдельные слова и знаки пунктуации.

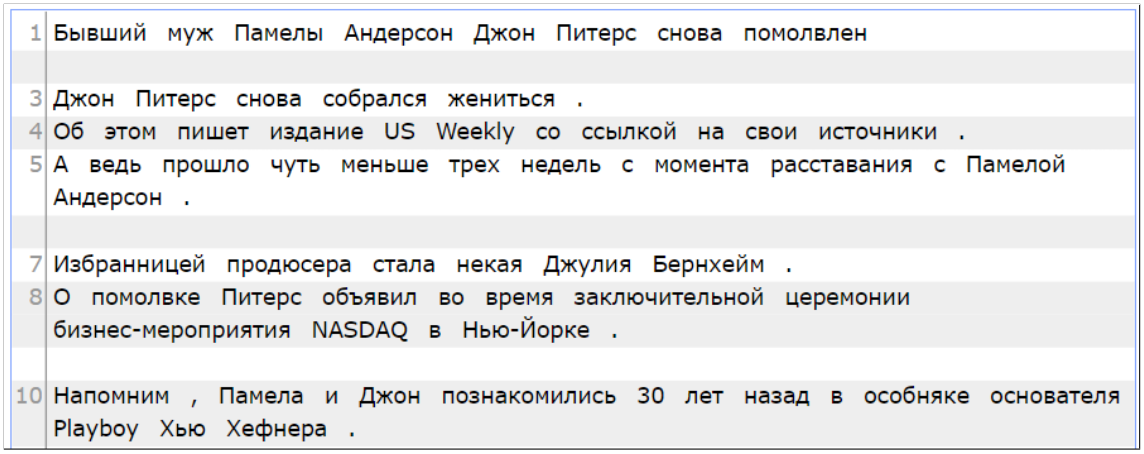

Рис. 21: Пример текста в интерфейсе `brat`

4.1 Уточняющие прилагательные

Прилагательные, уточняющие/конкретизирующие сущность (чаще всего они относятся к Person Property) входят в эту сущность, например: [российский президент], [чеченский блогер], [голливудский актер], [первый/бывший/будущий президент], [питерский "Зенит"], [младшая дочь],и т.п.

Обычные качественные прилагательные, типа *красивый, хороший, большой, дорогой* и т. п., которые дополнительно не конкретизируют объект, не входят в сущность.

4.2 Предлоги

Начальные предлоги могут входить в ряд типов сущностей:

- Все типы сущностей групп **Date** и **Time**
- Все типы сущностей группы **Numeric**, кроме Ordinal

Примеры предлогов в составе сущностей изображены на Рисунке 22.

Примечание. В конец сущности могут входить уточняющие предлоги, например: "`[в трёх километрах от]`$_{Quantity}$ `деревни`", "`[за два часа до]`$_{Time:Relative}$ начала".

4.3 Знаки препинания

Знаки препинания входят в сущность, только если они употребляются в середине сущности (запятые в адресах, дефисы в номерах телефонов, двоеточия и т. п.).

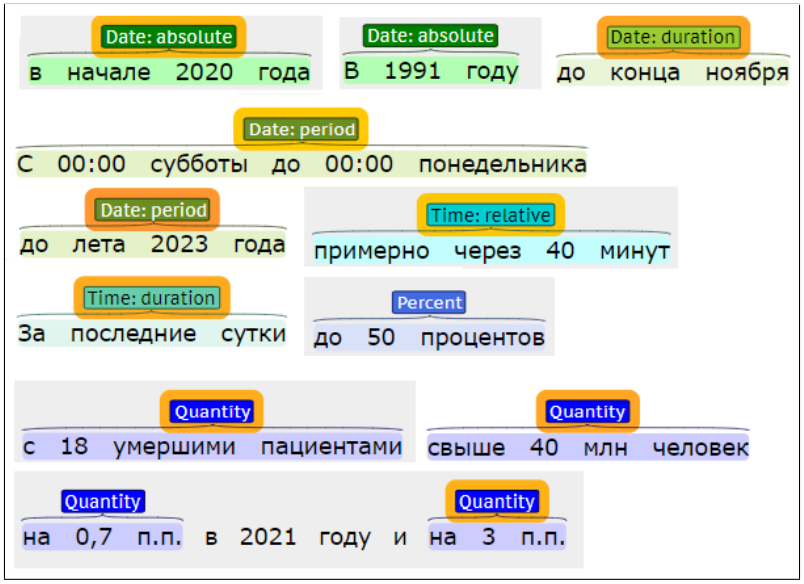

Рис. 22: Предлоги в составе сущностей

Конечные знаки препинания могут входить в сущность только если они являются частью названия объекта. Часто это восклицательные знаки, как, например, в *Mama Mia!* и *Кто боится Вирджинии Вульф?* (названия фильмов).

Запятые, точки, восклицательные, вопросительные и другие знаки препинания, не входящие в названия, не являются частью сущности (Рисунок 23).

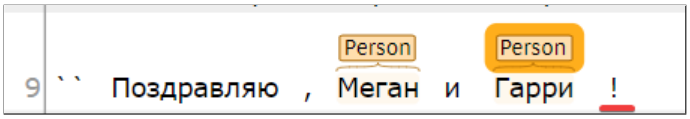

Рис. 23: Знаки препинания, как правило, не входят в сущность

4.4 Кавычки

В текстах начальные кавычки отображаются как ``, закрывающие кавычки - ".

Кавычки входят в сущность, если обособляют названия объектов (Рисунок 24). Кавычки, обозначающие начало/конец прямой речи или цитаты в сущности не входят (Рисунок 25).

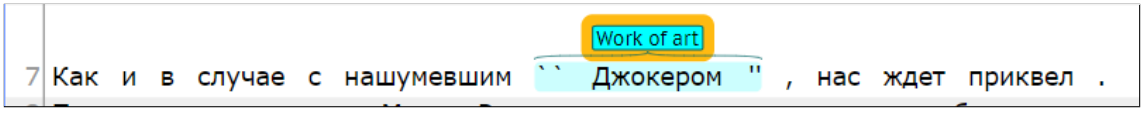

Рис. 24: Пример кавычек в названиях. Кавычки входят в сущность.

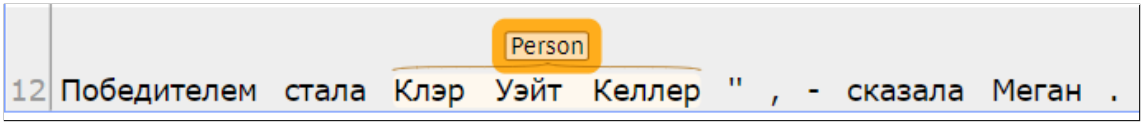

Рис. 25: Пример кавычек в прямой речи. Кавычки не входят в сущность.

4.5 Временные выражения с неопределенными границами

Выражение, обозначающее неопределенное время или временной промежуток, выделяется как сущность, только если оно несет смысловую нагрузку. Примеры таких выражений - *сейчас, в настоящее время, в настоящий момент, на данный момент* и др. Как правило, они обозначают неопределенный временной промежуток относительно времени написания статьи. Если они являются важными для понимания смысла контекста, их следует разметить как **Time/Date** → **Relative**.

Тест на наличие смысловой нагрузки временного выражения. Уберите временное выражение из предложения. Если смысл предложения не изменился, временно выражение не несет смысловой нагрузки и не отмечается как сущность (Таблица 1).

Таблица 1: Примеры временных выражений с неопределенными границами.

Выражение	**Имеет смысловую нагрузку**	**Не имеет смысловой нагрузки**
сейчас	*Сейчас я еще студент, а завтра уже буду выпускником.*	*Сейчас я пекарь. = Я пекарь.*
в настоящий момент	*В настоящий момент суд принимает решение об оправдании, мы узнаем решение через пару часов.*	*В настоящий момент на Земле проживает 7,5 млрд человек. = На Земле проживает 7,5 млрд человек.*

4.6 Одиночные именованные сущности и анафора

Анафоричные упоминания именованных сущностей, не имеющие смысла без антецедента, не выделяются как именованная сущность.

Например: *Белорусская оппозиция* *раскрыла план действий при отказе Лукашенко покинуть пост. Если требования [оппозиции] не будут выполнены, последует ряд массовых забастовок.*

В примере выше только выражение "*Белорусская оппозиция*" выделяется как именованная сущность. Отдельно упомянутое слово "*оппозиция*" не является именованной сущностью, так как без антецедента нельзя установить, какая именно это оппозиция.

При разметке кореференции выражение "*оппозиция*" следует выделить как анафор.

Исключением из правила является выражения "*столица*", "*граница*" и "*правительство*". Если в тексте подразумевается столица и Правительство

России, “*столица*” размечается как именованная сущность *GPE*, “*граница*” – как *Facility*, а “*правительство*” – как *Organization*.

Примечание. Одиночные слова, не указывающие на конкретный объект, не выделяются как сущность. К таким, например, относятся *полиция, армия, власти, маркетплейсы, соцсети, ТВ, радио*.

4.7 Numeric: Quantity vs. Cardinal в случае эллипсиса

Единичные числа могут отмечаться как Quantity, если в тексте объект счета пропущен, но может быть восстановлен из контекста. На Рисунке 26 выделенное выражение можно восстановить до “*умер 7071 человек*”.

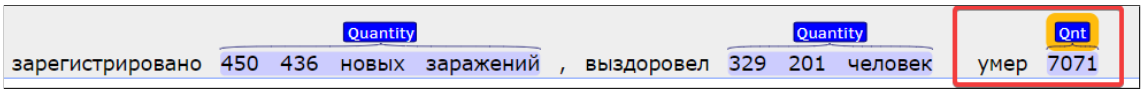

Рис. 26: Пример единичного числа в роли Quantity

4.8 Выделение сущности, разбитой на две строки

Если в результате сегментации текста одна сущность оказалась на нескольких строках, необходимо сначала выделить часть сущности на первой строке, выбрать подходящий тип сущности и нажать `Enter` или OK. Затем снова зайдите в выбор типа сущности, нажмите на Add Frag. и выделите оставшуюся часть сущности на второй строке (Рисунок 27).

После выделения фрагмента на второй строке, появится связь между двумя фрагментами (Рисунок 28).

Примечание. Если выделить разбитую на две строки сущность целиком, в интерфейсе появится ошибка (см. Раздел 4.9).

4.9 Возможные ошибки и предупреждения

В некоторых случаях могут появляться ошибки или сообщения с предупреждениями. Возможные случаи:

- **Пересекающиеся сущности**. Наложение сущностей друг на друга недопустимо. В случае пересечений метки подсвечиваются красным (Рисунок 29). Необходимо переразметить сущности таким образом, чтобы они не пересекались.

- **Пробелы в начале сущности.** Если в начало выделенной сущности попали пробелы, высветится предупреждающее сообщение (Рисунок 30). Чтобы предупреждение пропало, необходимо переразметить сущность (с

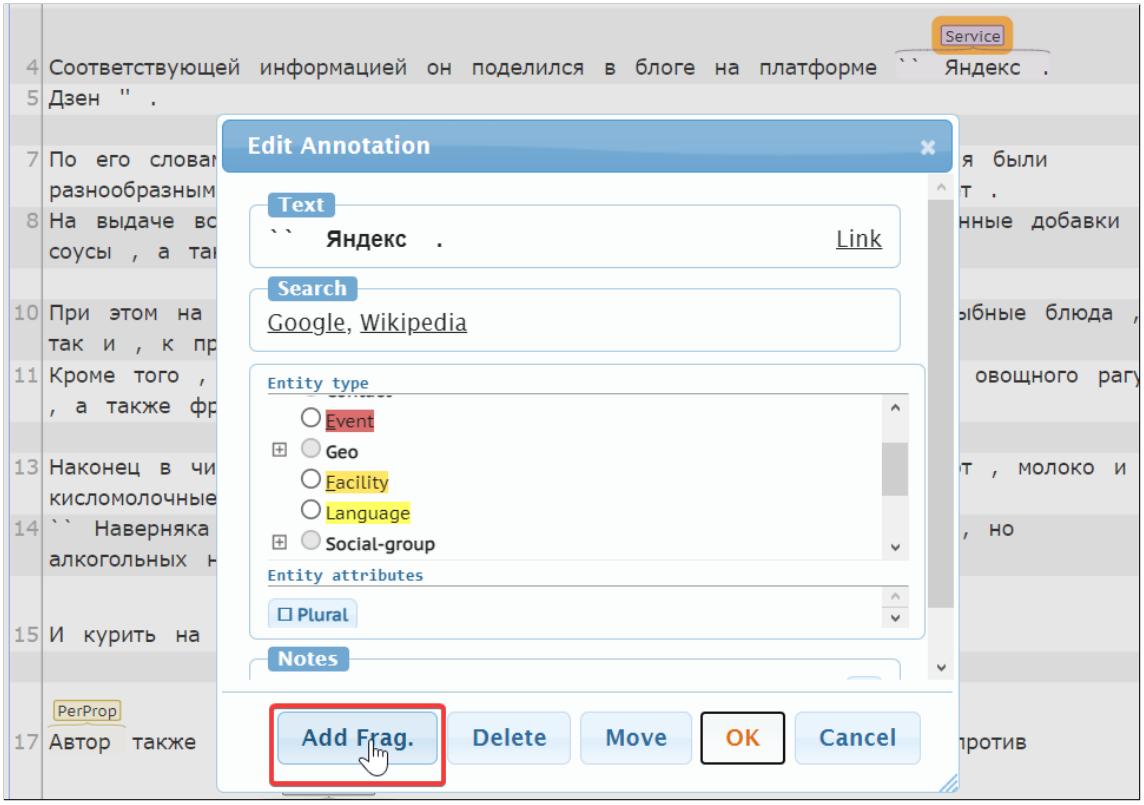

Рис. 27: Добавление фрагмента к сущности

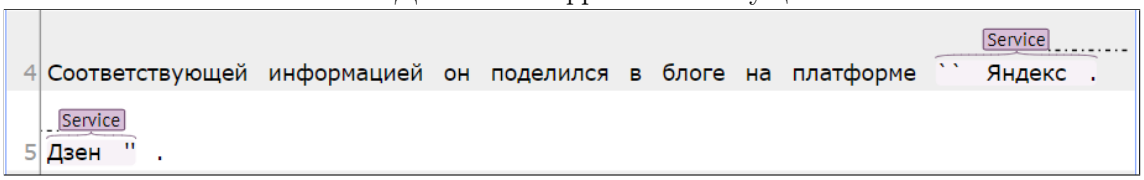

Рис. 28: Связь между фрагментами

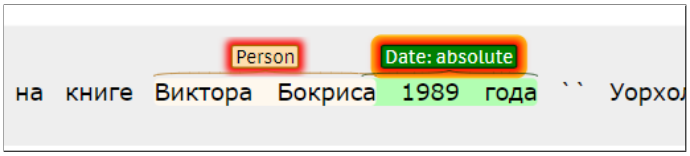

Рис. 29: Пример ошибки при пересечении сущностей.

помощью **Move** или **Delete**), при выделении не включая лишние пробелы в начало.

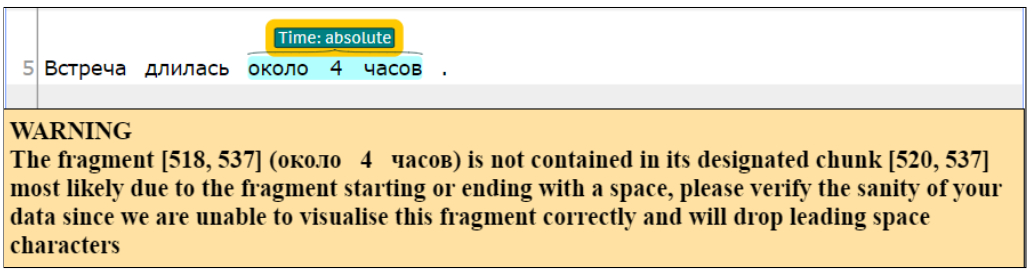

Рис. 30: Пример предупреждения при наличии пробелов в начале сущности.

Примечание. Лишние пробелы в конце сущностей будут удалены автоматически, они не вызывают сообщений с предупреждениями или ошибок.

- **Случайное выделение части второй строки при выделении сущности** вызовет серию ошибок, как изображено на Рисунке 31. Удалить такую сущность самостоятельно и убрать ошибку не получится, обратитесь к администратору. В разделе 4.8 описано, как выделить сущность, разбитую на две строки, так, чтобы не возникало ошибок.

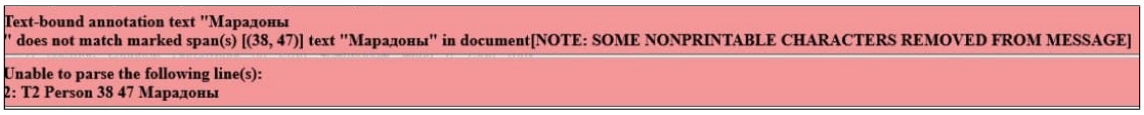

Рис. 31: Пример ошибки при выделении второй строки.

- **Ошибка сегментации.** Могут встречаться тексты, в которых цельные предложения разбиты на несколько строк (например, как на Рисунке 28). Если на границе строк одна сущность разбита на фрагменты, необходимо разметить ее, как описано в разделе 4.8. **Редко:** Если деление предложения не проходит по именованной сущности, в первой строке разбитого предложения отметьте любое слово, не являющееся сущностью, как !!!INVALID DOCUMENT!!! и напишите комментарий ‘‘`Ошибка сегментации`’’. При этом остальной текст следует разметить как обычно.

A Changelog

11.01.2021

1. **Обновлена Таблица 1**: Добавлен атрибут `Plural` в тип сущности `Event`. Добавлены новые примеры в некоторые типы сущностей. Удалены противоречивые примеры из категории `Person`. Сущность `Event` перемещена выше в таблице и в интерфейсе `brat`.

2. **Добавлен подраздел [DELETED]** *Выражения с упоминанием GPE*, в т.ч. Таблица [DELETED] с примерами.

3. **Добавлен раздел 4.5** *Временные выражения с неопределенными границами*, в т.ч. Таблица 1.

12.01.2021

1. **Обновлена Таблица 1**: Добавлен атрибут `Plural` в сущности `GPE` и `Location`. Удалены ошибочные примеры из категории `Facility`. В описание `Facility` добавлены “*улицы, площади, переулки*”. Добавлены примеры в категорию `Language`, удалены противоречивые примеры.

2. **Добавлен подраздел [UPGRADED TO SUBSECTION]** *Различия подтипов сущностей Time и Date*, в т. ч. Рисунок 16 с наглядной диаграммой различий подтипов временных сущностей.

3. **Добавлен раздел 4.6** *Именованные сущности и анафора.*

13.01.2021

1. **Добавлен раздел 4.8** *Выделение сущности, разбитой на две строки.* Добавлена ссылка на этот раздел в пункте *Случайное выделение...* в разделе 4.9.

2. **В разделе 4.9 добавлен пункт** *Ошибка сегментации.*

14.01.2021

1. **Обновлена Таблица 1.** В описание категории `Work of art` добавлены "*фильмы, ТВ шоу, сериалы*". Добавлены уточняющие вопросы в описание некоторых категорий группы `Numeric`, отредактировано описание `Quantity`. В категорию `Person` добавлен атрибут `Unconscious`.

2. **В разделе 2.3** добавлено описание нового атрибута `Unconscious`.

15-17.01.2021

1. **Обновлена Таблица 1.**

   - В описание `Event` добавлено уточнение "*не одиночные слова типа "концерт", "церемония" и т. д.*".
   - Изменен порядок сущностей в группе `Numeric`, обновлены их Hotkey.
   - Добавлены типы сущностей `Numeric:Age` и `Social-group:Family`.
   - В `Quantity` согласованы и добавлены примеры "*дважды, вчетверо реже, ни разу, однажды (в значении "однократно, один раз")*' и т. п.
   - В `Other group` добавлены примеры "*избиратели, бюджетники, поклонники, эксперты*".
   - В `Event` добавлены "президентские выборы в США".
   - Исправлено описание категорий группы `Date` и `Time` (уточнение про кванты/изменения).
   - В тип сущности `Product` добавлены атрибуты `Project` и `Trademark`.

- Чтобы уместить таблицу с новыми изменениями, уменьшен размер шрифта примеров.

2. **Добавлен Раздел 2.2** *Вложенные и пересекающиеся сущности*, в т.ч. ряд скриншотов с примерами.

3. **Добавлен подраздел 4.1** *Уточняющие прилагательные.*

4. **В разделе 2.3** добавлены новые атрибуты `Project` и `Trademark`.

5. **Удален раздел "Пояснения к некоторым типам сущностей"**. Ввиду разрешенной вложенности сущностей, он больше не актуален, вся необходимая информация по типам сущностей - в колонке *Описание* Таблицы 1. Под-подаздел *Различия подтипов сущностей Time и Date* выделен в отдельный подраздел 2.4.

6. **Удален ряд неактуальных или избыточных скриншотов.**

7. **Расширен Раздел 4.6**. Добавлена информация о выделении или невыделении одиночных слов как именованные сущности по итогу обсуждения с fostroll.

8. Кроме вышеперечисленного, были добавлены незначительные небольшие изменения в некоторые разделы в соответствии с новым допущением вложенности и подправлены примеры в Таблице 1.