

ru:corner

Разметка именованных сущностей и кореференции в brat

Анастасия Никифорова	Сергей Терновых
<code>steysie@gmail.com</code>	<code>fostroll@gmail.com</code>

Денис Киреев	Константин Ремизов
<code>dkireev.71@gmail.com</code>	<code>mr.enslin@mail.ru</code>

Январь 2021

Содержание

1	Общая информация о разметке в brat	1
1.1	Начало работы. Авторизация	1
1.2	Выбор текста из коллекции	1
1.3	Выделение и разметка текста	2
1.3.1	Алиасы сущностей в интерфейсе	3
1.3.2	Выделение сущности, разбитой на две строки	4
1.3.3	Комментарии к сущностям	4
1.3.4	Изменение и удаление меток	5
1.4	Кореферентные связи между упоминаниями	5
1.5	Нерелевантные тексты и ошибки токенизации	6
1.6	Поиск незнакомых терминов в Google и Википедии	6
1.7	Поиск по документу и коллекции	7
2	Разметка именованных сущностей в brat	13
2.1	Вложенные и пересекающиеся сущности	13
2.1.1	Вложенность или отдельные сущности?	14
2.1.2	Про одиночные Facility, Event и Service	14
2.2	Атрибуты типов сущностей	15
2.3	Общие правила разметки именованных сущностей	16
2.3.1	Уточняющие прилагательные	16
2.3.2	Уточняющие именные группы	17
2.3.3	Уточняющие предлоги и наречия	17
2.3.4	Знаки препинания	17
2.3.5	Кавычки	17
2.4	Особенности разметки некоторых типов сущностей	18
2.4.1	Person с вложенными Person Name, Person Property и др.	18
2.4.2	Family с Person и Person Property	20
2.4.3	GPE и Loc. Атрибут adj	23
2.4.4	Выбор между Service, Service:Media, Org и Org:Media	23
2.4.5	Product: продукт, серия, уникальное название – как быть?	24
2.4.6	Numeric: Quantity. Пересечения Quantity	25
2.5	Разбор Time, Date и Duration	25
2.5.1	Подтипы Time	26
2.5.2	Подтипы Date	26
2.5.3	Date в Event и наоборот	29
2.5.4	Ordinal в Date	30

2.6	Одиночные именованные сущности и анафора	30
3	Разметка кореференции в brat	31
4	Возможные ошибки и предупреждения	33
	Приложение A Changelog	35

1 Общая информация о разметке 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.

1.1 Начало работы. Авторизация

Разметка в **brat** доступна только для авторизованных пользователей. Чтобы авторизоваться, наведите курсор к шапке страницы, над текстом. В появившейся строке выберите **Login** (Рисунок 1).

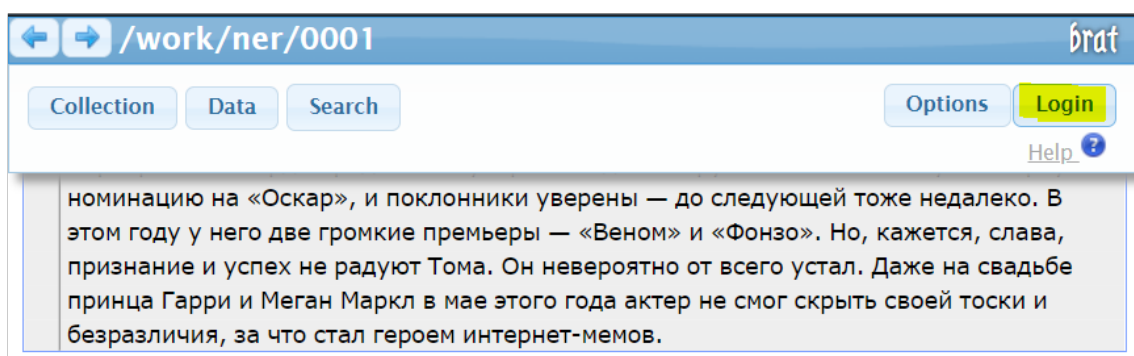


Рис. 1: Авторизация в brat

В появившемся окне введите логин и пароль и нажмите **ОК** (Рисунок 2). После успешной авторизации внизу страницы появится приветственное сообщение.

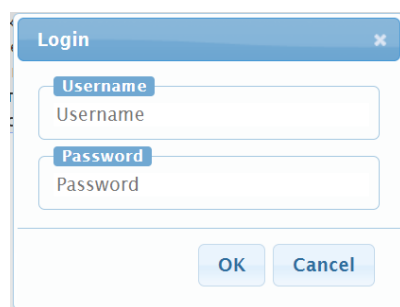


Рис. 2: Окно авторизации

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

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 3). Откроется окно с текстом как на Рисунке 4.

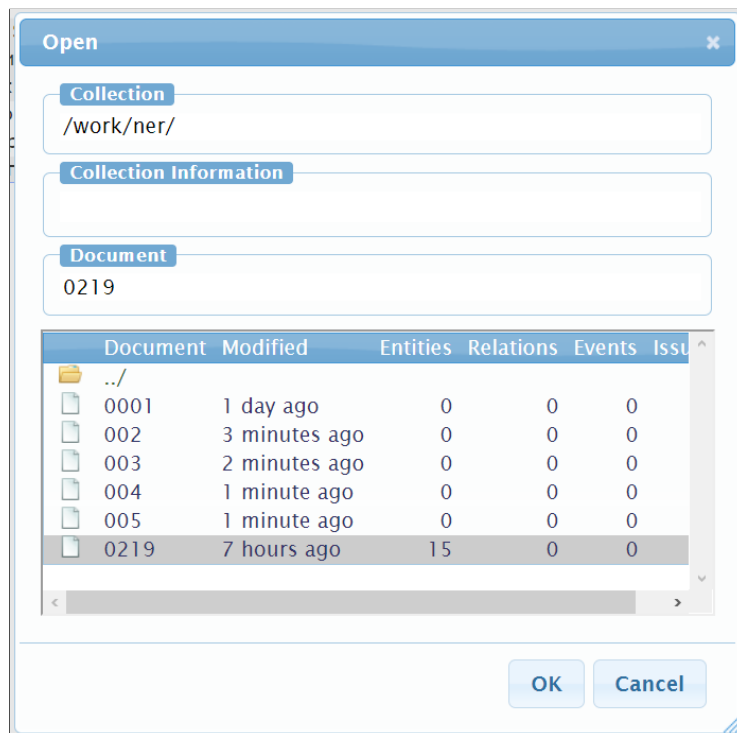


Рис. 3: Пример окна Collections

1.3 Выделение и разметка текста

В интерфейсе **brat** между всеми токенами¹ для удобства разметки – три пробела. Каждое предложение отображается на новой строке, а между каждым абзацем есть пустая строка. Пример текста изображен на Рисунке 4.

1	Бывший муж Памелы Андерсон Джон Питерс снова помолвлен
3	Джон Питерс снова собрался жениться .
4	Об этом пишет издание US Weekly со ссылкой на свои источники .
5	А ведь прошло чуть меньше трех недель с момента расставания с Памелой Андерсон .
7	Избранницей продюсера стала некая Джулия Бернхейм .

Рис. 4: Пример текста в интерфейсе brat

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.

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

¹Токены – отдельные слова и знаки пунктуации.

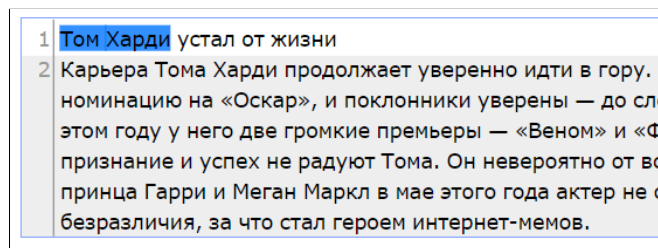



Рис. 5: Пример выделенного отрезка текста

Рис. 6: Пример окна выбора метки именованной сущности

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

Когда весь текст размечен, нажмите клавишу **→** или кликните на кнопку  в левом верхнем углу **brat** для перехода к следующему тексту.

1.3.1 Алиасы сущностей в интерфейсе

В зависимости от длины выделенной фразы, в интерфейсе могут использоваться алиасы – укороченные варианты одного и того же типа сущности, – как показано на Рисунке 7.

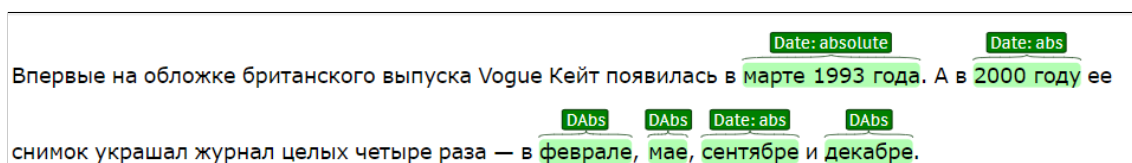


Рис. 7: Примеры алиасов сущности Date: absolute

1.3.2 Выделение сущности, разбитой на две строки

Если в результате сегментации текста одна сущность оказалась на нескольких строках, можно разметить сущность выделив ее целиком на всех строках. В таком случае, сущность автоматически разобьется на несколько фрагментов, как показано на Рисунке 9.

Если необходимо добавить фрагменты вручную, нужно сначала разметить часть сущности на первой строке, затем снова зайти в выбор типа сущности, нажать на **Add Frag.** (“Добавить фрагмент”) и выделить оставшуюся часть сущности на второй строке (Рисунок 8). Если строка разбита на большее количество строк – проделать предыдущий шаг с оставшимися строками.

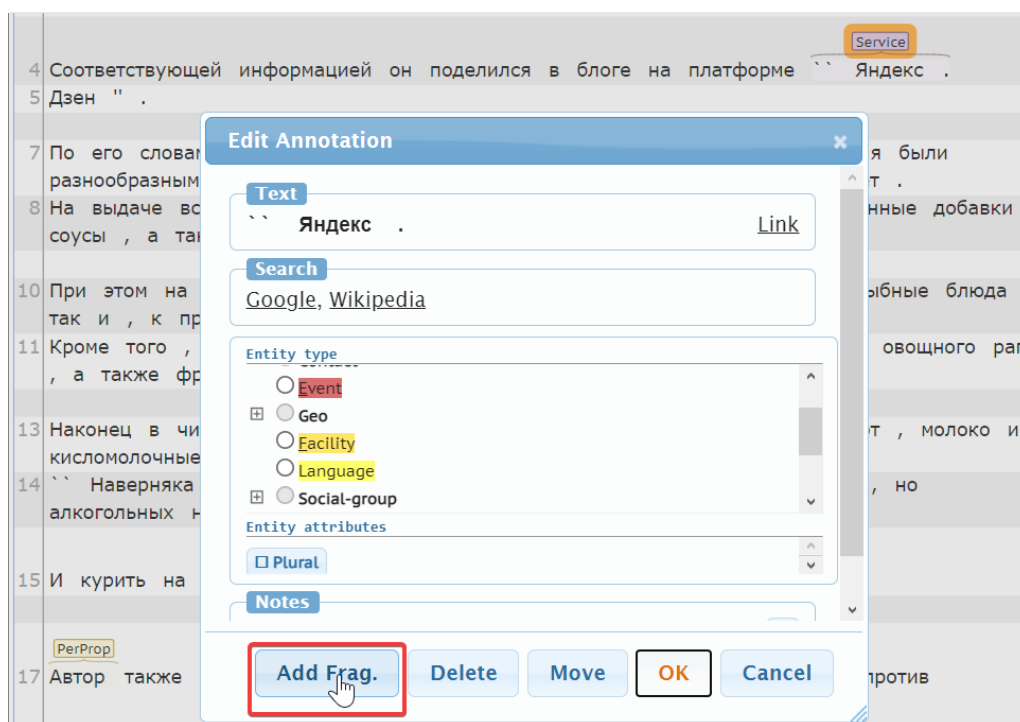


Рис. 8: Добавление фрагмента к сущности

После выделения фрагмента на второй строке, появится связь между двумя фрагментами одной сущности (Рисунок 9).

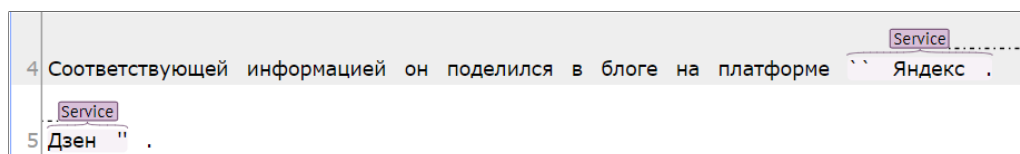


Рис. 9: Пример сущности из нескольких фрагментов

1.3.3 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**.

1.3.4 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).

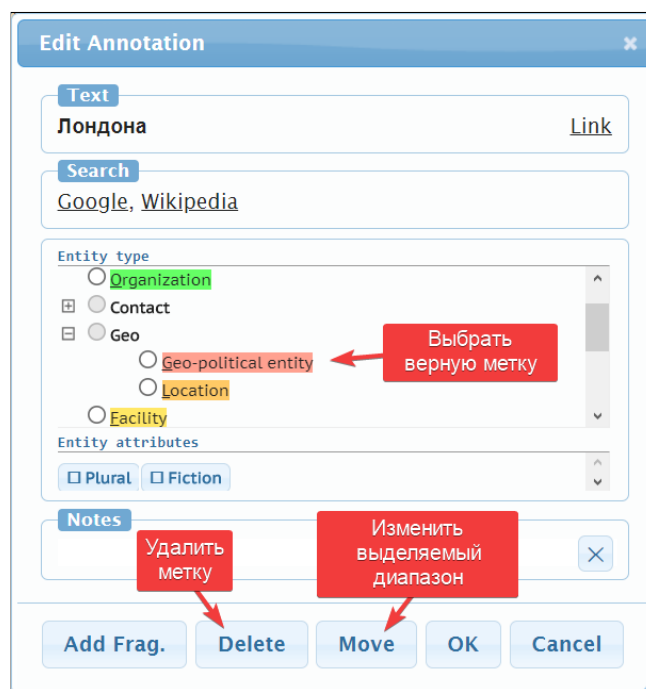


Рис. 10: Изменение и удаление меток именованных сущностей

Для изменения типа сущности, выберите другую метку из списка.

Для изменения границ сущности, нажмите **Move** в окне разметки или **Insert** на клавиатуре и заново выделите нужный диапазон. Во время изменения границ сущности рамка вокруг текста становится красной.

Для удаления метки, нажмите **Delete** в окне разметки или на клавиатуре.

1.4 Кореферентные связи между 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 11 и перетащите появившуюся стрелку на второе упоминание, как на Рисунке 12.

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 13).

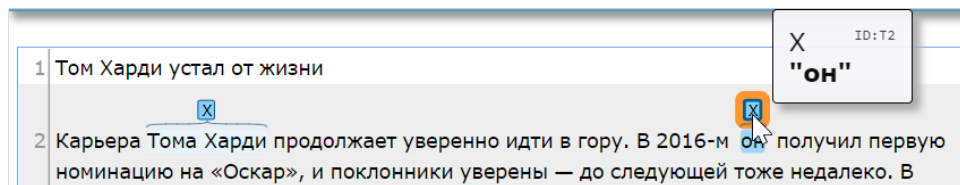


Рис. 11: Шаг 1. Выбор метки одного из упоминаний

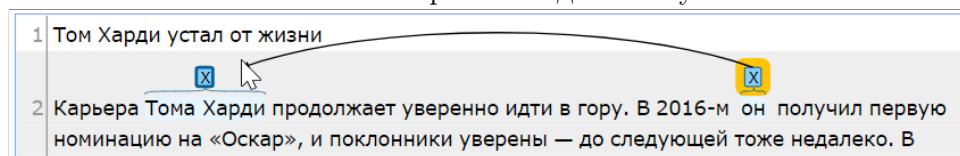


Рис. 12: Шаг 2. Связывание кореферентной группы

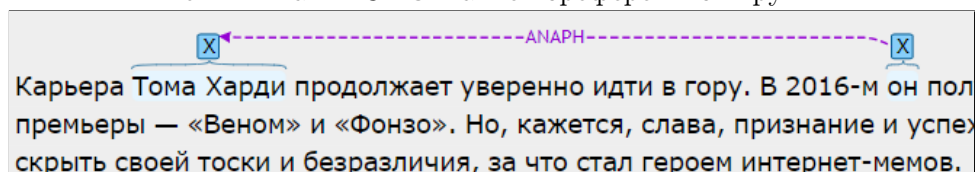


Рис. 13: Шаг 3. Выбор связи

1.5 Нерелевантные тексты и ошибки токенизации

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 первое слово специальным тегом `!!! INVALID DOCUMENT !!!`. Весь остальной текст следует оставить неразмеченным, как показано на Рисунке 14.

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

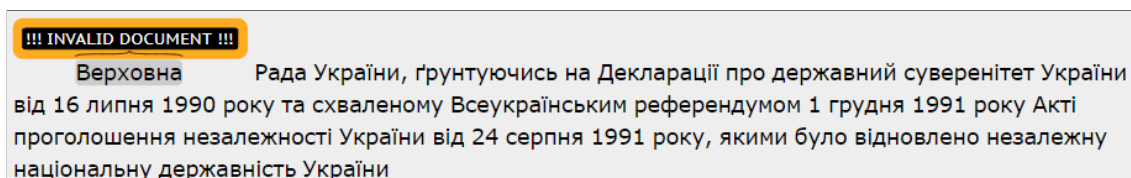


Рис. 14: Пример разметки нерелевантного текста

Кроме того, тегом `!!! INVALID DOCUMENT !!!` следует помечать отдельные случаи ошибок токенизации – когда токены не разделены на отдельные, как на Рисунке 15. При этом в склеенном токене следует выделить возможные сущности.

1.6 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 16).

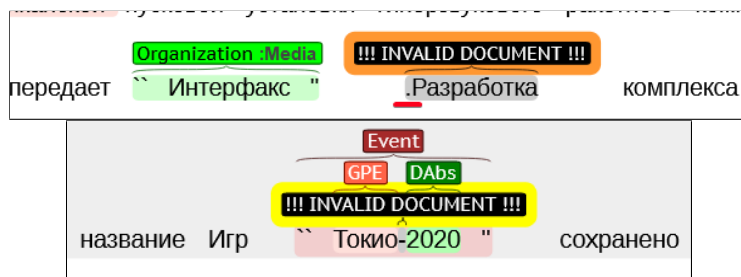


Рис. 15: Пример разметки ошибки токенизации

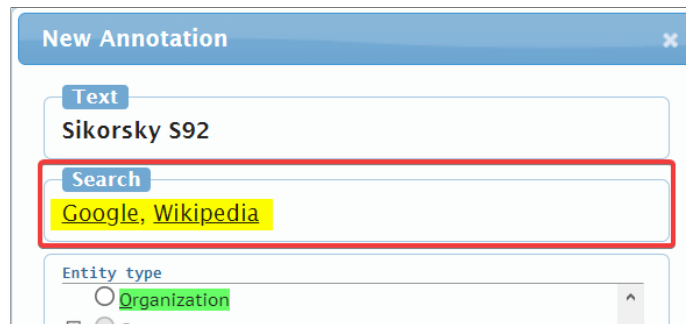


Рис. 16: Поиск незнакомых терминов в Google и Wikipedia

1.7 Поиск по документу и коллекции

В brat предусмотрен удобный поиск по тексту и коллекции.

Для вызова окна поиска, нажмите **Ctrl+F** или **Search** в левом верхнем углу интерфейса. Расширенные параметры поиска появляются при нажатии **Show Advanced**.

На вкладке **Text** (Рисунок 17) можно искать любые вхождения нужной строки в тексте. На вкладке **Entity** (Рисунок 18) можно ограничить поиск: будут найдены только те строки, которые в коллекции входят в сущность (любую или конкретно заданного типа).

На Рисунках 17 и 18 изображены вкладки **Text** и **Entity** в окне поиска с описанием расширенных параметров поиска.

Для поиска² введите в поле *Text* фразу, слово, часть слова или регулярное выражение. В окне **Entity** можно вместе или вместо слова выбрать тип искомой сущности. Далее задайте необходимые параметры и нажмите **OK**.

В появившемся окне выберите первый документ – искомое слово в нем будет подсвечиваться оранжевым. Для прохода по всем найденным случаям, нажимайте на клавиатуре кнопку вправо → (обратно - влево ←), пока не дойдете до последнего элемента.

Пока работает поиск, коллекция отфильтрована по документам, в которых есть искомые слова. Для выхода из режима поиска, нажмите на крестик на

²Помните, что в текстах ru_corner три пробела между словами. Если нужно найти фразу – между каждым токеном должно быть три пробела.

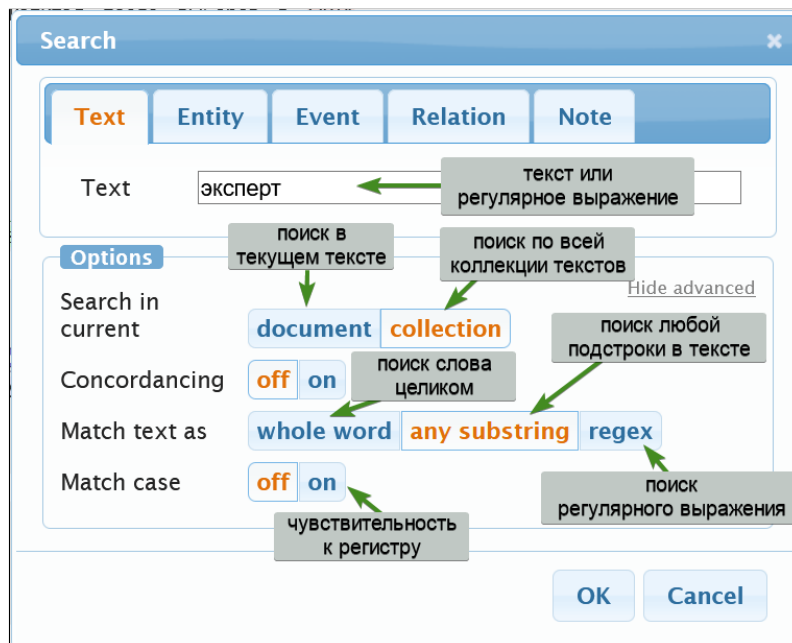


Рис. 17: Окно поиска. Вкладка Text



Рис. 18: Окно поиска. Вкладка Entity

кнопке **Search** в левом верхнем углу.

Поиском рекомендуется пользоваться после разметки всей коллекции для проверки возможных пропусков и ошибок.

Примечание. Опция **Concordancing: on** в расширенном меню позволяет вывести заданное количество символов до и после каждой из найденных подходящих строк. С помощью этой опции можно быстро найти нужную фразу в нужном документе, изучив окно с результатами поиска.

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		<input type="checkbox"/> Plural <input type="checkbox"/> Fiction <input type="checkbox"/> Unconscious 0 или 1 из: <input type="checkbox"/> Male (муж) <input type="checkbox"/> Female (жен)	P	Полное 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
		<input checked="" type="checkbox"/> Plural		Имена нескольких людей в одной сущности	Павел и Марина Смирновы, Дэвид и Виктория Бэкхам
		<input checked="" type="checkbox"/> 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, кот Матроскин, Мальчик который выжил, Тот-чьё-имя-нельзя-называть
		<input checked="" type="checkbox"/> Unconscious		Любые имена, которыми люди называют неразумные уникальные объекты: животные, техника, корабли и т. п. (включая тип объекта, если он указан, напр., "корабль")	Барсик, Белка, Стрелка, Несси (лох-несское чудовище), броненосец "Потемкин", корабль "Сметливый", баркас "Fishizzle", "Союз МС-16", эсминец Джон Маккейн, корабль Орион, ледокол "Арктика"
Person Name	Forename		1	Имя человека, включая второе и последующее имена	Михаил, Мишаня, Майкл, Николь, Франсуа, "Хиллари Дайан Родэм"
	Surname		2	Фамилия	Иванов, Путин, Клинтон, Дикаприо, да Винчи, ван де Херик
	Patronym			Отчество (в разных культурах разные конструкции, при затруднении - искать в Google)	Петрович, Семеновна, Юрьич, ибн Мухамед, Тимер улы, Бусурманкул Уулу, бен Рашид
	Initial		I	Инициал	И., Т., О., Дж.
	Alias		A	Псевдоним	Мариванна, Бейонсе, Моргенштерн, Грозный, аль-Заяни
	Complex name			Сложное имя, которое не получается разделить на составные части	"Аль-Малик ан-Насир Салах ад-Дунийя ва-д-Дин Абуль-Музаффар Юсуф ибн Айюб ибн Шади аль-Курди"
	Affix			Префиксы и суффиксы имени	Мистер, мисс, ее высочество, сквайр, старший, младший
Person property		<input type="checkbox"/> Plural <input type="checkbox"/> Fiction	R	Чин, звание, титул, должность, профессия и т. п.	генерал-майор, королева, канцлер, сварщик, менеджер, президент, ученые, разработчики, депутаты, чиновник, папа римский, исследователи, террорист, специалист, эксперт, заключенные, задержанные, пенсионер
		<input checked="" type="checkbox"/> Plural		Несколько Property в одной сущности	"главы России, Азейбарджана и Армении"
		<input checked="" type="checkbox"/> Fiction		Вымышленные титулы/профессии и т. п.	профессор защиты от темных искусств, штурмовик Первого Ордена
Organization		<input type="checkbox"/> Plural <input type="checkbox"/> Fiction <input type="checkbox"/> Department <input type="checkbox"/> Media	O	Компании, агентства, радиостанции, институты, политические партии, армии стран и т. п., имеющие орг. структуру	Ростелеком, Amazon, ПАО "Газпром", партия "Единая Россия", РПЦ, компания Орифлейм, Белорусская оппозиция, турецкая армия
		<input checked="" type="checkbox"/> Plural		Несколько организаций или отделов в рамках одной сущности	Норильский и Алтайский горнодобывающие заводы, финансовый и юридический отделы, Московская и Екатеринбургская епархии
		<input checked="" type="checkbox"/> Fiction		Вымышленные организации (кино, литература и т. п.)	галактический сенат, Stark Industries, Снектр, школа волшебства Хогвартс
		<input checked="" type="checkbox"/> Department		Отделы внутри организаций	IT-департамент, отдел кадров, совет директоров, топ-менеджмент
		<input checked="" type="checkbox"/> Media		СМИ: издательства, журналы, газеты и т.п., включая печатные издания	издательство ЭКСМО, Эхо Москвы, Лента.ру, Cosmopolitan, АиФ, Первый канал
Contact	Address			Адрес местоположения, зачастую - город, улица, дом, квартира, индекс	123456, Москва, Тверская улица, дом 10, строение 2; адрес: ул. Ленина, д. 5
	Phone			Номер телефона	+7 (123) 456-78-90, тел. 81234567890, телефон 44-22-33
	Email			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

Таблица 1. Типы сущностей, их описания и примеры

Категория	Тип сущности	Атрибуты	Hotkey	Описание типа сущности	Пример
Event		<input type="checkbox"/> Plural	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, пандемия коронавируса, эпидемия, предизидентские выборы в США, бой Хабиба и Коннора, концерт Баскова и Киркорова (один концерт)
		<input checked="" type="checkbox"/> Plural		Несколько Event в рамках одной сущности	ураганы Катрина и Рита, концерты Баскова и Киркорова (разные мероприятия)
Geo	Geo-political entity (GPE)	<input type="checkbox"/> Plural <input type="checkbox"/> Fiction <input type="checkbox"/> Adjective	G	Географическая зона, имеющая политическую структуру + космические станции	Ростов-на-Дону, Швейцария, Ближний Восток, г. Москва, СНГ, СССР, Катманду, деревня Кунцево, Алтунфьевский район, столица (если понятно, что Москва), МКС
		<input checked="" type="checkbox"/> Plural		Несколько GPE	Московская и Ленинградская области
		<input checked="" type="checkbox"/> Fiction		Вымышленная географическая зона	Атлантида, Хогвартс, Нарния, Вестерос, Средиземье, Готэм-сити, Асгард
		<input checked="" type="checkbox"/> Adjective		GPE, выраженные прилагательными	омский, чешский, замбийский, российский, московский,
	Location	<input type="checkbox"/> Plural <input type="checkbox"/> Fiction <input type="checkbox"/> Adjective	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, естественный спутник Земли, космос
		<input checked="" type="checkbox"/> Plural		Несколько Location	реки Тигр и Евфрат
		<input checked="" type="checkbox"/> Fiction		Вымышленные места	Мглистые горы, Андуин, река Яруга
		<input checked="" type="checkbox"/> Adjective		Location, выраженные прилагательными	черноморский, тихоокеанский, гималайский
Social-group	Nationality	<input type="checkbox"/> Citizenship <input type="checkbox"/> Plural <input type="checkbox"/> Resident <input type="checkbox"/> Adjective	H	Принадлежность к нации по происхождению, рождению или иным образом	русские, итальянец, грек, папуасы, финны, афроамериканцы, кореец, индус, армянского происхождения
		<input checked="" type="checkbox"/> Plural		Несколько разных объектов Nationality в одной сущности	Граждане Грузии и Армении, жители Северной и Южной Кореи
		<input checked="" type="checkbox"/> Citizenship		Гражданство какой-либо страны	граждане РФ, гражданка Эстонии, жители Китая, американцы
		<input checked="" type="checkbox"/> Resident		Жители городов, областей, провинций, в том числе фразы "местные жители", "жители страны" и т. п.	москвичи, жительница Стерлитамакского района Башкирии, жительница канадской провинции Онтарио, жители страны, жители России, местные жители
		<input checked="" type="checkbox"/> Adjective		Nationality, выраженное прилагательным. При выборе Adjective, другие оставить пустыми.	русский, татарский, турецкий, еврейский, арабский, израильский
	Family		Y	Обозначения родственных связей	мать, брат, старший сын, сводная сестра, двоюродная тетя, брат жены, семья, род, сожители, пара, молодожены
	Religious group	<input type="checkbox"/> Plural <input type="checkbox"/> Adjective <input type="checkbox"/> Source (src)		Принадлежность к определенной религии	православные, католик, старообрядцы, амиши, шииты, атеисты, пастафарианцы
		<input checked="" type="checkbox"/> Plural		Несколько разных религиозных групп в одной сущности	православные, католические и протестантские христиане
		<input checked="" type="checkbox"/> Adjective		Религия, выраженная прилагательным	христианский, католический, мусульманский
		<input checked="" type="checkbox"/> Source (src)		Религиозное направление	христианство, буддизм, агностицизм, язычество, ислам, конфуцианство
	Political group	<input type="checkbox"/> Plural <input type="checkbox"/> Adjective <input type="checkbox"/> Source (src)		Принадлежность к политической группе	республиканцы, члены партии "Единая Россия", национал-демократы, коммунисты, члены партии «Свобода»
		<input checked="" type="checkbox"/> Plural		Несколько политических групп в одной сущности	члены КПРФ и ЛДПР, члены партий Единая Россия и Яблоко
		<input checked="" type="checkbox"/> Adjective		Политическое направление, прилагательное	демократический, либеральный, коммунистический
		<input checked="" type="checkbox"/> Source (src)		Политическая идеология	коммунизм, демократия, анархизм, консерватизм, либерализм, национализм, социализм
	Other group	<input type="checkbox"/> Plural <input type="checkbox"/> Fiction		Другие социальные группы, члены которых имеют что-то общее	безработные, веганы, гангстеры, новые русские, готы, геи, бомжи, преступник, избиратели, бюджетники, поклонники
		<input checked="" type="checkbox"/> Plural		Несколько разных групп в одной сущности	представители готов и эмо, участники клубов рукоделия и гончарства
		<input checked="" type="checkbox"/> Fiction		Вымышленные социальные группы	хоббиты, эльфы, орки, джедаи, члены банды "Железные рукава"

Таблица 1. Типы сущностей, их описания и примеры

Категория	Тип сущности	Атрибуты	Hotkey	Описание типа сущности	Пример
Facility		<input type="checkbox"/> Plural <input type="checkbox"/> Fiction	F	Достопримечательности, здания, аэропорты, шоссе, парки, мосты, улицы, площади, переулки и т. п.	памятник Ильичу, Внуково, Трасса М4, ТРЦ "Авиапарк", Дворцовый мост, Арбат, Трехсвятская улица, проспект Ленина
		<input checked="" type="checkbox"/> Plural		Несколько Facility	трассы М4 и М5
		<input checked="" type="checkbox"/> Fiction		Вымышленные Facility (кино, литература и т. п.)	Замок Саурона, Замок Дарта Вейдера, Галактический рынок
Language		<input type="checkbox"/> Plural <input type="checkbox"/> Fiction		Языки естественные или искусственные (не из кино/литературы)	русский язык, итальянский, иврит, хинди, язык йоруба, урду, эсперанто, северные диалекты русского языка
		<input checked="" type="checkbox"/> Plural		Несколько объектов Language в рамках одной сущности	южные и восточные диалекты, тюркские и финноугорские языки, говоры и наречия русского языка
		<input checked="" type="checkbox"/> Fiction		Вымышленные языки (кино, литература)	клингон, на'ви, дотракийский, новояз
Product		<input type="checkbox"/> Plural <input type="checkbox"/> Project <input type="checkbox"/> Trademark	U	Электроника, автомобили, оружие, продукты питания, одежда, техника, серии водных/космических судов и т. п. (конкретные продукты с указанием серии/производителя)	пылесос Dyson, диваны Ikea, Range Rover, Айфон, Орион чокопай, пистолет Макарова, АК-47, косметика Орифлейм, платье бренда Виктория Бекхам, танк Т34, космические корабли "Союз", Катюша (оружие)
		<input checked="" type="checkbox"/> Plural		Несколько продуктовых сущностей в рамках одной	диваны Ikea и Hoff, пылесосы Dyson и Xiaomi, косметика Эивон и Орифлейм
		<input checked="" type="checkbox"/> Project		Проекты, программы (государственные, научные, космические и т. п.)	мегасаенс-проект NICA, нацпроект "Наука", программа обмена студентами Erasmus +, программа «Вояджер»
		<input checked="" type="checkbox"/> Trademark		Бренд, запатентованная технология, торговая марка продукта (не в значении "компания/организация")	Dyson, Apple, Орифлейм, Виктория Бекхам, Шанель, Tesla, технология ProMotion
Service		<input type="checkbox"/> Plural <input type="checkbox"/> Media	S	Различные предприятия, приложения и технологии предоставления услуг, онлайн-платформы	чистка обуви, ремонт ноутбуков, доставка еды, скорая помощь, Яндекс.Еда, Telegram, TikTok, YouTube, Facebook
		<input checked="" type="checkbox"/> Plural		Несколько сервисов в одной сущности	услуги прачечной и химчистки, услуги телефонии и интернета
		<input checked="" type="checkbox"/> Media		Аккаунты, каналы, страницы, сайты, сообщества и т. п. (главный признак - они могут публиковать контент и имеют уникальное имя)	Телеграмм-канал "IT юмор", страница Ольги Бузовой, сайт ООН, твиттер-аккаунт Трампа, в т. ч. одиночные "пост", "страница" и др., если есть уточнение платформы
Work of art				Названия книг, песен, картин, фильмов, ТВ шоу, сериалов и т. п.	Властелин колец, Мона Лиза, "Белые розы", Черный квадрат Малевича, Вечерний Ургант
Law		<input type="checkbox"/> Plural		Нормативно-правовые акты: законы, статьи (без названий, если есть номер статьи) и т. п.	Конституция, статья 20.6.1 КоАП, УК РФ, пункт "а" статьи 105, статья "Об умышленном убийстве", Закон Южной Кореи о защите информации, Закон о защите прав потребителей, антикоррупционное законодательство
		<input checked="" type="checkbox"/> Plural		Несколько Law в одной сущности	п. 2 статьи 228 и пп. 3 и 4 статьи 230 УК РФ
Date	Date: absolute	<input type="checkbox"/> Plural обязат-но 1 из <input type="checkbox"/> Past <input type="checkbox"/> Present <input type="checkbox"/> Future	D	Явная дата или конструкция, которую можно заменить на дату (когда? какой день/месяц/год?), включая "следующий/текущий/нынешний/прошедший" и т.п.	20 октября 2000 г., январь 2021 года, 2020, в марте, "в среду, 2 марта", 1 мая прошлого года, на прошлой неделе, ранее в четверг, утром 9 ноября, сегодня, сегодня рано утром, в настоящее время, во время саммита
	Date: relative	<input type="checkbox"/> Past <input type="checkbox"/> Past-UNK <input type="checkbox"/> Past-Pres <input type="checkbox"/> Past-Fut <input type="checkbox"/> Pres-Fut <input type="checkbox"/> UNK-Fut	Z	Относительная дата, которая зависит от реального события, к которому относится	после случившегося, до этого, после которого, впоследствии, ранее, позднее, через несколько часов, спустя два года, на 7-м месяце беременности
	Date: period absolute	<input type="checkbox"/> Past <input type="checkbox"/> Past-UNK <input type="checkbox"/> Past-Pres <input type="checkbox"/> Past-Fut <input type="checkbox"/> Pres-Fut <input type="checkbox"/> UNK-Fut	X	Реальный период времени. Атрибуты указывают на начало и конец периода	с 5 января по 10 февраля, за сентябрь, за последние шесть лет, на майские праздники, с 15 ноября до Нового года, отныне, "на два месяца, до 8 января 2021 года"
	Date: period relative	<input type="checkbox"/> Fut <input type="checkbox"/> LCont <input type="checkbox"/> RCont	C	Реальный период времени относительно другого события/даты (зависит от контекста)	в течение двух дней, за семь дней, чуть более года, два часа подряд,

Таблица 1. Типы сущностей, их описания и примеры

Категория	Тип сущности	Атрибуты	Hotkey	Описание типа сущности	Пример
Time	Time: absolute	<input type="checkbox"/> Plural	T	Время, когда что-то происходит однажды, не весь период (когда? во сколько?)	15:30, полтретьего, пятнадцать минут одиннадцатого, около 13:30, утром, днем, вечером, ночью, полночь, полдень, в ночь на, 00:00 по местному времени
	Time: relative	<input type="checkbox"/> Plural	V	Слова, указывающие на порядок действий или событий	сперва, затем, в то же время, одновременно, потом, дальше, после
	Time: period		B	Период времени, абстрактный или гипотетический	в год, за два года, за сутки, 10-20 дней после, за четыре дня
Duration			N	Продолжительность времени (как долго?)	три часа, сутки, день, два дня, месяц, год, пять лет, долгое время, несколько часов, многолетний, не первый день, третий месяц
Numeric	Ordinal		3	Числа и слова для подсчета предметов (какой по счету?)	первый, второй, тысячный, следующий, предыдущий, (пред)последний, повторно, первичный, впервые, несколько
	Money		4	Денежная сумма, включая название валюты, в т. ч. "от ... до ..." (сколько денег?)	\$500, два евро, 100 рублей, один тенге, 300 динаров, 55 франков, 45 рублей 22 копейки, от 20 до 30 тыс. руб., от \$200 до \$300, более 2 млн рублей, почти \$100, около 50 евро
	Percent		5	Проценты (включая "%") и дроби, в т. ч. слово "половина"	десять процентов, полпроцента, 99.9%, 147%, на 30.1% меньше, половина, треть, две трети, 1/3, 25-я часть, десятые доли процента
	Age		6	Возраст человека или предмета	25 лет, 20-летний, годовалый, молодой, юный, старше 65 лет, молодежь, ребенок, маленькие дети, от 3 до 17 лет, несовершеннолетний, новорожденный, подросток,
	Quantity		8	Измерения, количество чего-либо с единицей измерения (сколько?), включая счета матчей и фразы с "несколько"	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 "один раз"), пара кексов, тройка лошадей, единственный, топ-10, 5:2, несколько
Other term		<input type="checkbox"/> Plural <input type="checkbox"/> Fiction	Q	Термины, не подпадающие под категории выше, включая научные термины, болезни, технологии и т. д.	гигабайт, рибонуклеотид, минорные актиноиды, премия "Оскар", COVID-19, коронавирус, инсульт, ишемия миокарда
		<input checked="" type="checkbox"/> Plural		Несколько разных терминов, объединенных в одну сущность	гепатит А и В, азотная и соляная кислоты
		<input checked="" type="checkbox"/> Fiction		Вымышленные термины	Скайнет, криптонит

2 Разметка именованных сущностей в brat

Для разметки именованных сущностей в `brat` предусмотрено более 50 различных типа сущностей. Подробное описание сущностей с примерами приведено ниже в Таблице 1.

Некоторые сущности имеют дополнительные возможные атрибуты. Подробнее об атрибутах – в Разделе 2.2.

Для наиболее частотных типов сущностей предусмотрены горячие клавиши (колонок 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.

Если помимо типа сущности необходимо отметить атрибуты, клавиша **Enter** сработает только если вы снова нажмете на Hotkey этой сущности после выбора атрибутов. Либо можно кликнуть мышкой на ОК.

2.1 Вложенные и пересекающиеся сущности

Именованные сущности могут быть составными, то есть могут включать в себя вложенные сущности (а-ля “матрёшка”), или быть пересекающимися с другими сущностями. Примеры разметки с пересечениями и вложенностью изображены на Рисунке 19.

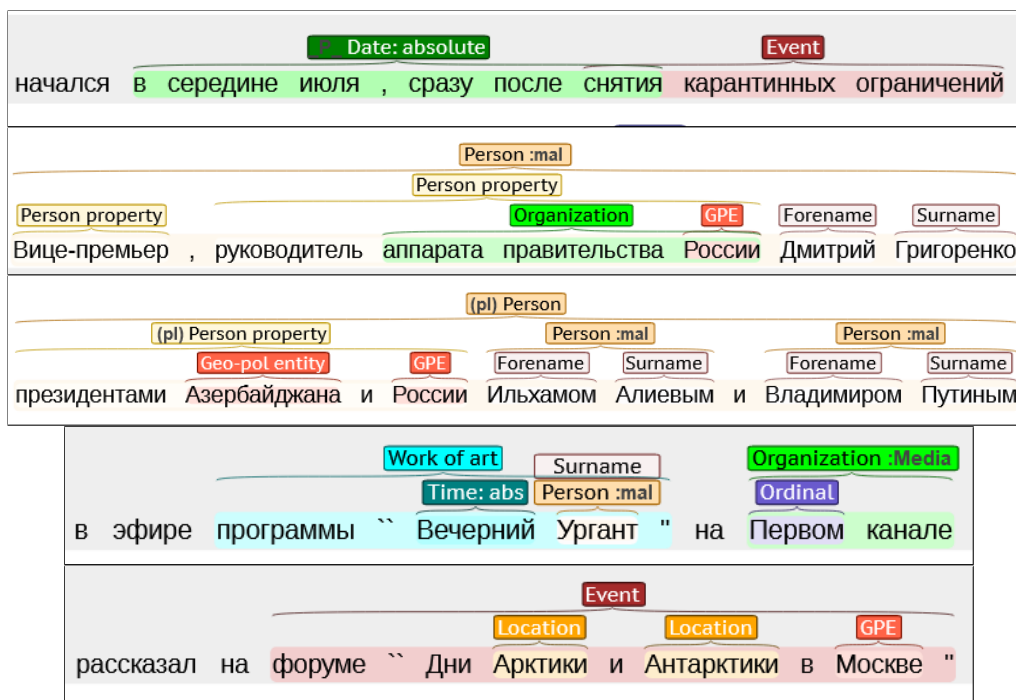


Рис. 19: Примеры пересекающихся и вложенных сущностей

Примечание. Пересекающимися могут быть только сущности из катего-

рии **Quantity** (количество), **Date: relative**, **Date: period absolute** и **Date: period relative**. Пример пересечения сущностей: “{шестерых [членов] экипажа}”. Можно ориентироваться на то, что пересечение неразрешенных типов сущностей подсвечивается красным. В таком случае, нужно использовать вложенность.

2.1.1 Вложенность или отдельные сущности?

Сущности могут содержать в себе вложенные сущности, с которыми они, как правило, связаны управлением³ и их не разделяют предлоги или другие слова (см. “Вложенность” в Таблице 2).

Если сущности разделены предлогом или другими словами, они выделяются отдельно, без вложенности (см. “Раздельные сущности” в Таблице 2).

В Таблице 2 приведены различия цельных, вложенных и отдельных сущностей.

Таблица 2: Примеры цельных, вложенных и отдельных сущностей

Одна сущность	Вложенность	Раздельные сущности
[воронежский аэропорт]	[аэропорт [Воронежа]]	[аэропорт] в [Воронеже]
[немецкий канцлер]	[канцлер [Германии]]	[канцлер] в [Германии]
***	[Бурзянский район [Республики Башкортостан]]	[Бурзянский район] в [Республике Башкортостан]
***	[село Старосубхангулово [Бурзянского района [Республики Башкортостан]]]	[село Старосубхангулово] в [Бурзянском районе] в [Республике Башкортостан]
***	[ЦК [партии “Коммунисты России”]]	[ЦК] в [партии “Коммунисты России”]
***	[бой [Головкина] против [Альвареса]]	между [Головкиным] и [Альваресом] состоялся [бой]

2.1.2 Про одиночные Facility, Event и Service

Одиночные сущности, относящиеся к категориям **Facility**, **Event** и **Service** (например, “аэропорт” и “бой” в Таблице 2, и упоминания сервисов “сайт”, “форум” и т.п.), могут выделяться как именованная сущность, только если в этом же абзаце в пределах трех предложений справа и слева, включая центральное, есть уточнение (местоположение, год и т. п.), конкретизирующее эту одиночную сущность (как “в Воронеж^е”, “между Головкиным и Альваресом” в Таблице 2; для сервисов – это упоминание организации, которой принадлежит сервис).

³**Управление** – вид подчинительной связи, при которой зависимое слово употребляется в том косвенном падеже, которого требует главное слово. В нашем случае – это, как правило, связь *существительное+существительное* (а точнее, *именная группа+именная группа*), например “[президент [Российской Федерации]]”

Примечание. Это правило не действует на одиночные упоминания, если в предыдущем контексте есть их полное название (например, [*Международный аэропорт Воронеж*] или [*сайт Всемирной организации здравоохранения*]) или если в этом абзаце в пределах трех предложений справа и слева, включая центральное предложение, нет уточнений/конкретизаторов этого упоминания.

2.2 Атрибуты типов сущностей

Некоторые сущности могут иметь один или несколько атрибутов. В Таблице 1 они перечислены в колонке **Атрибуты**. Как и сущности, выбор атрибутов зависит от контекста: в разных контекстах одна и та же сущность может иметь разные атрибуты или не иметь атрибутов. Ниже приведен список основных атрибутов и их значений в алфавитном порядке.

- **Adjective (adj)** – сущность в форме прилагательного. Доступно для **GPE** (*московский, российский*), **Location** (*тихоокеанский, алтайский*), **Nationality** (*русский, арабский*), **Political group** (*либеральный, коммунистический*) и **Religious group** (*христианский, католический*).
- **Citizenship** – гражданство какой-либо страны (*американец, поляки*). Доступно только для сущности **Nationality**.
- **Department** – отдел внутри организации. Доступно только для **Organization**.
- **Fiction** – вымышленная сущность, как правило, относящаяся к литературе, кино, сериалам и др.
- **Male (mal)** / **Female (fem)** – мужской или женский пол сущности **Person**.
- **Media** – любые медиа-источники: издательства, социальные сети, телевизионные каналы и т. п. Доступно для **Organization** и **Service**.
- **Plural** – два или более разных объекта, объединенных в одну именованную сущность как “*объекты X и Y*”.
- **Project** – проекты, программы (государственные, научные, космические и т. п.). Доступно только для типа сущности **Product**.
- **Resident** – житель какой-либо страны/города/населенного пункта. Доступно только для **Nationality**.
- **Source (src)** – в **Political group** и **Religious group** политическая идеология или религиозное направление соответственно.

- **Trademark** – бренд, торговая марка продукта (не в значении *Organization*). Доступно только для типа сущности **Product**.
- **Unconscious** – имена, которыми называют уникальные “неразумные” объекты: корабли, ракеты, животных и др. Доступно только для **Person**.

The screenshot shows a web form for entity classification. At the top, there's a 'Text' field containing 'издание Daily Mail' and a 'Link' button. Below that is a 'Search' field with 'Google, Wikipedia'. The 'Entity type' section features a list of radio buttons: 'Person', 'Person property', 'Organization' (which is selected and highlighted in green), 'Contact', 'Event', 'Geo', and 'Facility'. At the bottom, the 'Entity attributes' section contains four checkboxes: 'Plural', 'Fiction', 'Department', and 'Media' (which is checked and highlighted in orange). A red rectangle highlights the 'Entity attributes' section.

Рис. 20: 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 20).

Кроме того, сущности группы **Date** имеют свои атрибуты, которые подробно рассмотрены в Разделе 2.5.

2.3 Общие правила разметки именованных сущностей

2.3.1 Уточняющие прилагательные

Прилагательные, уточняющие/конкретизирующие сущность (чаще всего они относятся к **Person Property**) входят в эту сущность, например:

российский премьер-министр, чеченский блогер, голливудский актер,
первый/бывший/будущий президент, питерский “Зенит”, младшая дочь

и т. п.

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

2.3.2 Уточняющие именные группы

Уточняющие именные группы (существительное с зависимыми словами, если они есть) при именованных сущностях входят в эту сущность, например:

город Москва, радиостанция Эхо Москвы, издательство ЭКСМО,
американская авиакомпания Delta Air Lines, футбольный клуб “Зенит”,
река Волга и т. п.

2.3.3 Уточняющие предлоги и наречия

Уточняющие предлоги и наречия могут входить в ряд типов сущностей:

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

Примеры уточняющих предлогов и наречий:

после 4 декабря, свыше 200 случаев, более 25 тысяч рублей, до 14 лет,
с 1 по 5 мая, на 30 процентов меньше и т. п.

Примечание. В конец сущности могут входить уточняющие предлоги, например: “[в трёх километрах от]_{Quantity} деревни”, “[за два часа до]_{Time:Relative} начала”, “[через два дня после]_{Date:Relative} начала”.

2.3.4 Знаки препинания

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

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

Запятые, точки, восклицательные, вопросительные и другие знаки препинания, не входящие в названия, не являются частью сущности (Рисунок 21).

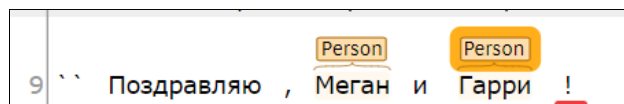


Рис. 21: Знаки препинания, как правило, не входят в сущность

2.3.5 Кавычки

Начальные кавычки отображаются как ``, закрывающие кавычки - ”.

Кавычки входят в сущность, если обособляют названия объектов (Рисунок 22). Кавычки, обозначающие начало/конец прямой речи или цитаты в сущности не входят (Рисунок 23).

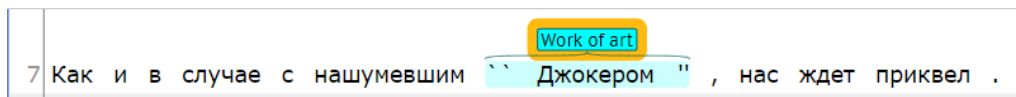


Рис. 22: Пример кавычек в названиях. Кавычки входят в сущность.

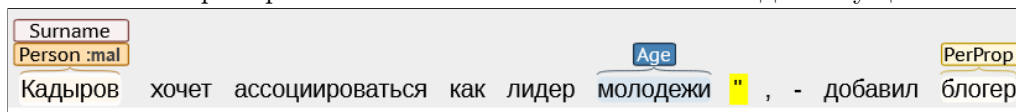


Рис. 23: Пример кавычек в прямой речи. Кавычки не входят в сущность.

2.4 Особенности разметки некоторых типов сущностей

В этом разделе описаны особенности разметки наиболее частотных типов сущностей.

2.4.1 Person с вложенными Person Name, Person Property и др.

Сущность **Person** – это живые или вымышленные люди, а также животные или названия уникальных единиц техники (корабли, ракеты и др.).

Person имеет несколько опциональных атрибутов (**Plural**, **Fiction**, **Unconscious**, описанные с Таблице 1), а также на выбор один из атрибутов **Male** (в интерфейсе - **mal**, мужской род) и **Female** (**fem**, женский род), если можно определить биологический пол лица, о котором идет речь.

Как правило, сущности **Person** всегда вложенные, поскольку даже в простых **Person**, выраженных именем, выделяются сущности **Person Name**⁴.

Важно! **Person Name** может существовать только внутри сущности **Person**.

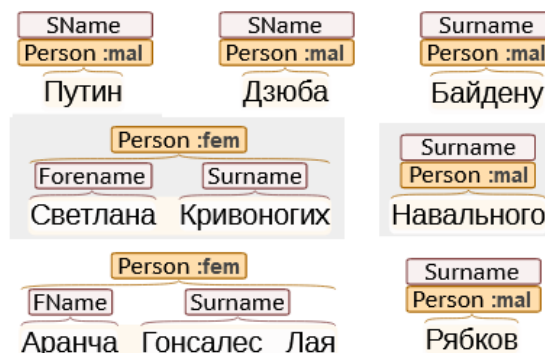


Рис. 24: Разметка одиночных имен: **Person+Person Name**

Person могут содержать в себе множество разных типов сущностей. Наиболее часто - **Person Name** и **Person Property**, а также другие **Person**. В свою очередь, вложенные в **Person** сущности могут содержать в себе другие вложенные сущности (Рисунки 24-25).

⁴За исключением некоторых редких случаев. Например, [корабль "Сметливый"] выделяется как **Person:Unconscious** без вложенных сущностей

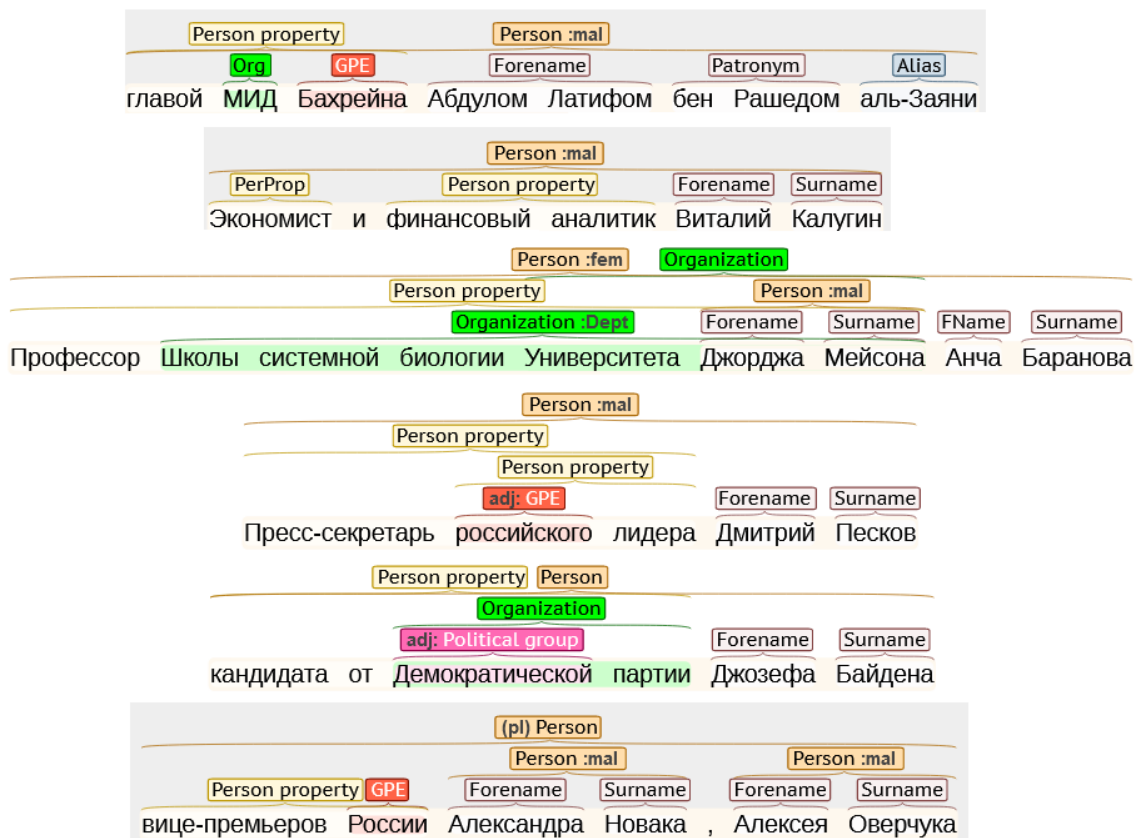


Рис. 25: Разметка вложенных Person

Исключением является сущность Age. Она не входит в сущность Person, если стоит перед ней (Рисунок 26). Если возраст указан в середине Person, он автоматически входит в Person (Рисунок 27).

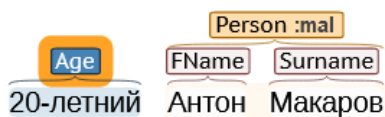


Рис. 26: Age перед Person

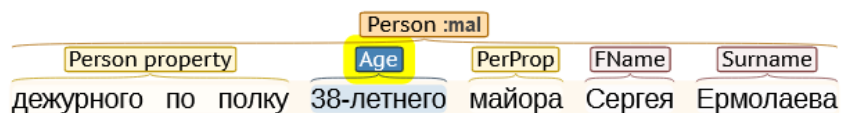


Рис. 27: Age внутри Person

Примечание. В основе Person может быть сущность Family. Подробнее о таких случаях в Разделе 2.4.2.

2.4.2 Family с Person и Person Property

Family – атомарная сущность, то есть в нее входит только термин, обозначающий родство (см. примеры в Таблице 1). У нее не может быть вложенных сущностей, однако она сама может быть вложенной сущностью.

В тексте будут встречаться разные комбинации сущности **Family** с **Person Property** и именами в именительном или родительном падеже⁵.

Кроме того, **Family** может встречаться вместе с притяжательными местоимениями и другими словами. Во всех таких случаях зачастую **Family** и зависимые слова объединяются в одну сущность **Person**.

Примечание. Родительный или именительный падеж у имени – это определяется при именительном падеже **Family**. Например “*сыну Борису*” → “*сын Борис*”; “*сына Бориса*” → в зависимости от контекста, “*сын Борис*” или “*сын Бориса*”.

Рассмотрим возможные сочетания и как их размечать.

Family + Имя в именительном падеже

Пример: “*брат Борис*”. Такое сочетание объединяется в одну сущность **Person**, при этом имя не является отдельной, вложенной сущностью **Person**, так как обозначает того же человека. Если перед **Family** есть притяжательное местоимение, оно включается в общую сущность **Person**. На Рисунке 28 приведены примеры разметки таких сочетаний.

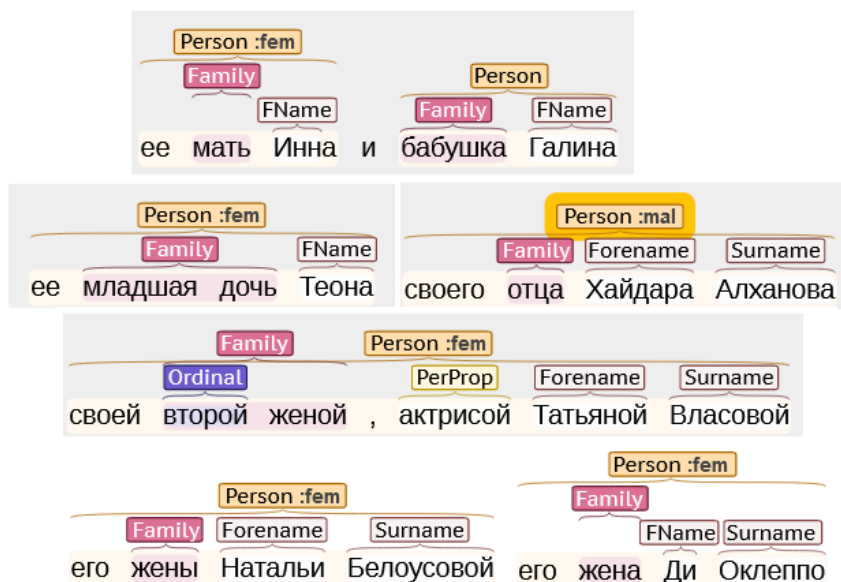


Рис. 28: Family + Имя в именительном падеже

⁵Сравните: “сын Иван” vs. “сын Ивана”. В первом примере чье-то сына зовут Иван, а во втором - у Ивана есть сын, чье имя не названо.

Family + Имя в родительном падеже

Пример: “*брат Бориса*”. Такое сочетание объединяется в одну сущность **Person**, при этом имя дополнительно размечается как **Person**, так как обозначает другого человека. Если перед **Family** есть притяжательное местоимение, оно включается во внешнюю сущность **Person**. На Рисунке 29 приведены примеры разметки таких сочетаний.

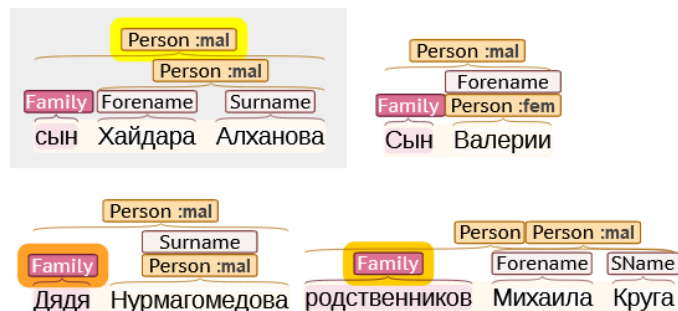


Рис. 29: Family + Имя в родительном падеже

Family + Имя в род. падеже + Имя в имен. падеже

Пример: “*брат Бориса Иван*”. Такое сочетание объединяется в одну сущность **Person**, при этом дополнительно как **Person** размечается только имя в родительном падеже, так как обозначает другого человека (чей родственник). Если перед **Family** есть притяжательное местоимение, оно включается во внешнюю сущность **Person**. На Рисунке 30 приведены примеры разметки таких сочетаний.

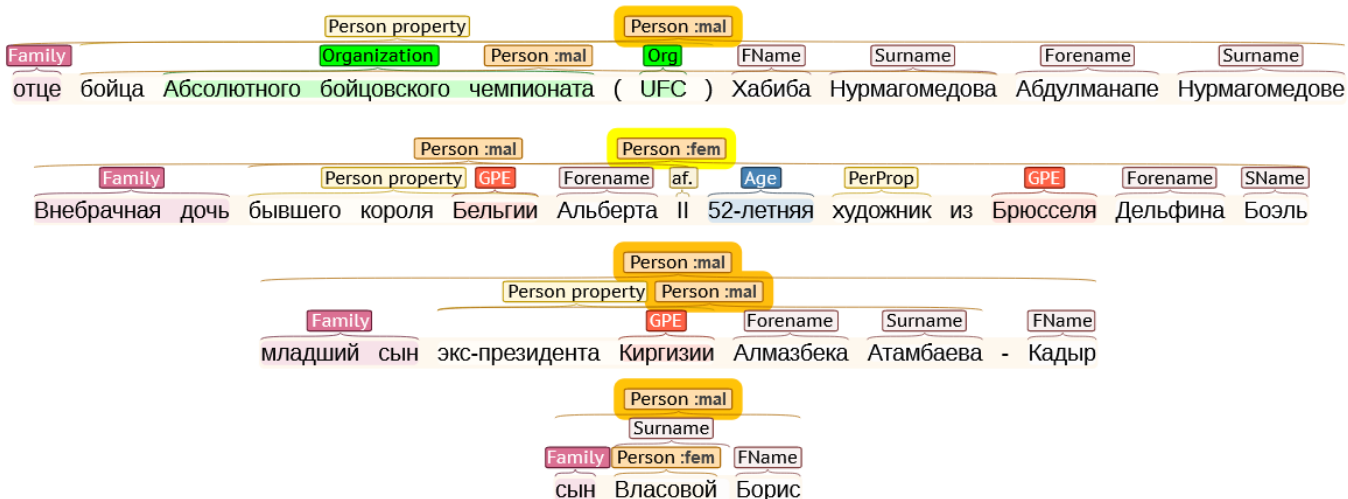


Рис. 30: Family + Имя в род. падеже + Имя в имен. падеже

Family + Property или похожее на Property слово в род. падеже

Пример: “*брат адвоката*”. Такое сочетание объединяется в одну сущность **Person**, при этом слово в родительном падеже размечается дополнительно,

только если оно подходит под определение **Person Property** или других категорий. Если перед **Family** есть притяжательное местоимение, оно включается во внешнюю сущность **Person**. На Рисунке 31 приведены примеры разметки таких сочетаний.



Рис. 31: Family + Property или похожее на Property слово в род. падеже

Притяжательное местоимение + Family

Пример: “его брат”. Такое сочетание объединяется в одну сущность **Person** (Рисунок 32). Исключениями являются слова, обозначающее цельное объединение родственников: *семья, род, династия* и т. п.

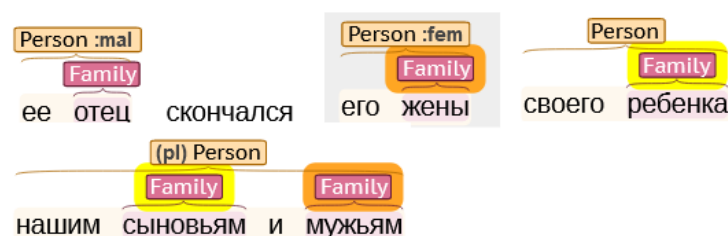


Рис. 32: Притяжательное местоимение + Family

Family и приложения/апозитивные конструкции

Пример: “Иван, брат Бориса, ушел.”. Если сущность **Family** с зависимыми словами (в род. падеже) встретилась в составе приложения или в качестве сказуемого в конструкциях типа “Иван – (это) сын Бориса”, она объединяется в сущность **Person** с зависимыми словами, отдельно от основного **Person**, как показано на Рисунке 33.

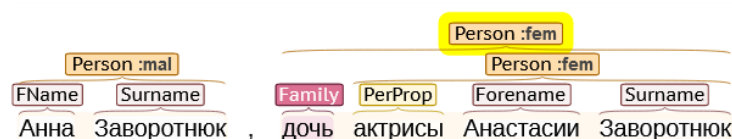


Рис. 33: Family и приложения/апозитивные конструкции

Исключения

Исключениями являются слова, обозначающие родство, но не относящиеся к людям, например, “*дочка компании Ростелеком*”. Также, не выделяются прилагательные типа “*отцовский, материнский, дочерний*” и т. д.

2.4.3 GPE и Loc. Атрибут adj

GPE – географические зоны, имеющие политическую структуру (мир, страны, города, районы и др.), а также космические станции.

Loc – природные места: горные цепи, водоемы, а также планеты, галактики, созвездия, кометы и т. п.

Для сущностей GPE и Loc доступен атрибут **Adjective** (в интерфейсе - **adj**, прилагательное). Этот атрибут выбирается, если сущность имеет значение характеристики объекта, его отнесённости к указанной местности (Рисунок 34).

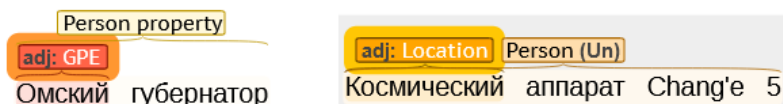


Рис. 34: Атрибут Adjective в GPE и Loc

Если имеется в виду сама местность, атрибут **Adjective** не ставится, даже если сущность выражена только прилагательным, как размечено на Рисунке 35. Уточняющее существительное типа *область, республика, село* может быть опущено, составные части Plural-сущностей также не размечаются как прилагательные.



Рис. 35: Случаи, когда атрибут Adjective не ставится

2.4.4 Выбор между Service, Service:Media, Org и Org:Media

При разметке упоминаний сайтов, соцсетей, страниц, каналов, аккаунтов и т. п. в соцсетях следует опираться на следующие примеры:

- Акции Facebook_{Org} подскочили

- Об этом написали в Комсомольской правде_{Org:Media}
- Об этом писали в социальной сети Facebook_{Service}
- Об этом писали в Facebook_{Service}
- Они написали об этом в своем Facebook_{Service:Media}
- Они написали обо этом на своей странице_{Service:Media} в Facebook_{Service}
- На канале_{Service:Media} Киркорова_{Person} в YouTube_{Service} вышел клип
- Об этом сообщают на сайте_{Service:Media} Кремля_{Org}
- Они занимаются раскруткой Instagram-аккаунтов_{Service:Media}
- Снимок был размещен на Telegam-канале “Записки охотника”_{Service:Media}

Примечание. Одиночные слова типа “*страница*”, “*канал*”, “*аккаунт*” размечаются как **Service:Media** только если в текущем абзаце есть уточнение, на какой платформе/сайте они находятся.

2.4.5 Product: продукт, серия, уникальное название – как быть?

В тексте могут встретиться разные комбинации из упоминаний типа продукта, серии/производителя, уникального названия. В Таблице 3 приведены примеры, как следует выделять разные сочетания этих упоминаний.

Таблица 3: Разметка разных сочетаний Product и Person:Unconscious

Сочетание	Как выделять	Пример
тип продукта & серия/производитель	Product (если внутри указан производитель или торговая марка - выделить вложенную сущность Product:Trademark)	[самолеты Су-57] [космические корабли серии “Союз”] [автомобиль [Форд]]
тип продукта & уникальное название	Person:Unconscious	[броненосец “Потёмкин”] [американский космический корабль Орион]
тип продукта & серия/производитель & уникальное название	[[тип продукта & серия] название] → [[Product] Person:Unconscious]	[[Ледокол типа ЛК60Я] “Арктика”]
тип продукта & уникальное название & серия/производитель	[[тип продукта & название] серия] → [[Person:Unconscious] Product]	[[Ледокол “Арктика”] типа ЛК60Я]
уникальное название	целиком Person:Unconscious	[Потемкин], [Орион]
серия	Product	[Катюша], [Су-57], [Т34]

Примечание. **Person:Unconscious** – уникальное название только сущностей **Product**. Уникальные названия сущностей **Event** не выделяются как **Person:Unconscious**. Следовательно, следующие примеры выделяются целиком как **Event** без вложенных сущностей: “Ураган Катрина”, “Циклон Сара”, “Музыкальный фестиваль Юрмала-93” и т. п.

2.4.6 Numeric: Quantity. Пересечения Quantity

Единичные числа могут отмечаться как **Quantity**, если в тексте объект счёта пропущен, но может быть восстановлен из контекста. Основной критерий **Quantity** – сущность должна отвечать на вопрос “сколько?”. На Рисунке 36 выделенное выражение можно восстановить до “умер 7071 человек”.

Сущности **Quantity** захватывают конкретизаторы (подробнее в Разделе 2.3.3), численное количество и ближайшую зависимую именную группу, обозначающую считаемые объекты. **Quantity** могут пересекаться с другими сущностями (Рисунок 37).

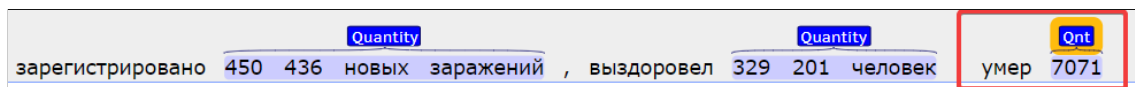


Рис. 36: Пример единичного числа в роли Quantity

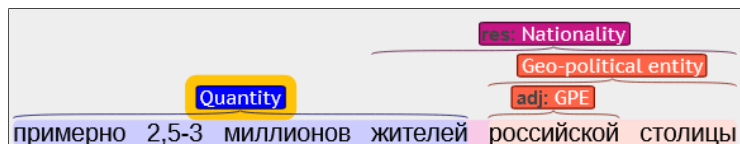


Рис. 37: Пример пересекающейся сущности Quantity

2.5 Разбор Time, Date и Duration

Для выражения времени используются три основных типа сущностей: **Duration**, **Date** и **Time**.

Duration – это продолжительность времени (примеры в Таблице 1). Этот тип сущности существует только внутри **Date** и **Time**, он не может быть размечен как самостоятельная сущность (Рисунок 38).

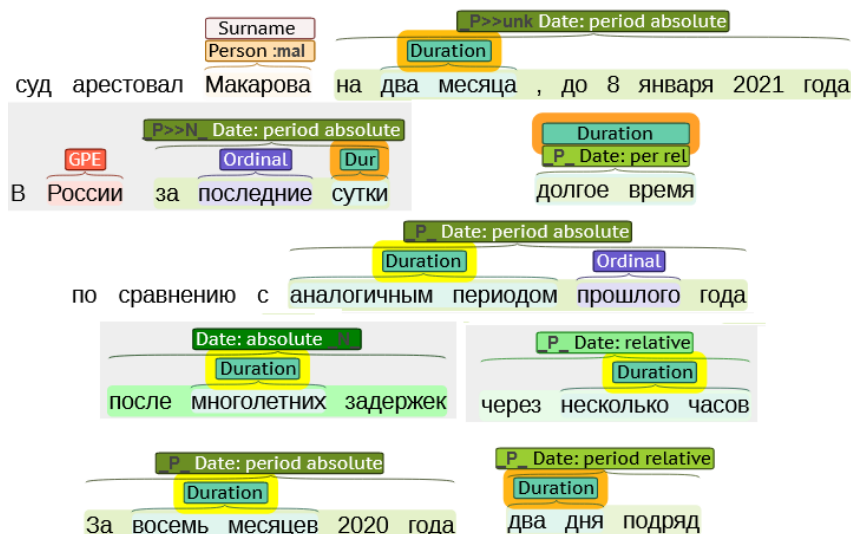


Рис. 38: Примеры разметки Duration

Time – это время, дату которого нельзя установить без контекста. В основном это либо обозначение времени в сутках, либо абстрактное или гипотетическое время. **Date** – это время, дату которого можно установить.

2.5.1 Подтипы Time

Time: absolute – время или период, когда что-то произошло/происходит/произойдет или может произойти конечное количество раз (не длится на протяжении всего периода). Отвечает на вопросы *когда?* *во сколько?* (Рисунок 39).

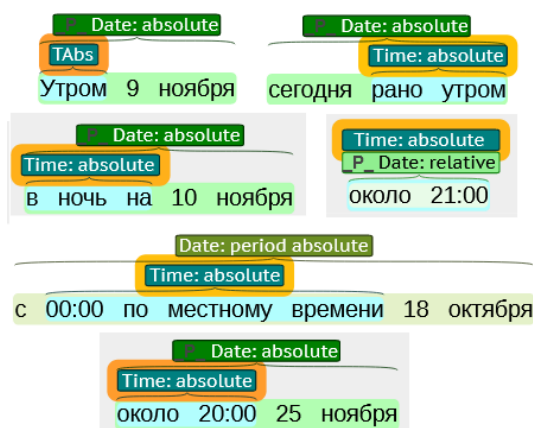


Рис. 39: Примеры разметки Time: absolute

Time: relative – слова, обозначающие порядок действий во времени относительно друг друга. Например, слова *сперва*, *затем*, *в то же время*, *одновременно*, *потом*, *дальше*, *после*.

Time: period – период времени: абстрактный или с неопределенными границами дат (Рисунок 40) или периодичность (Рисунок 41).



Рис. 40: Time: period – абстрактный период времени

2.5.2 Подтипы Date

Date: absolute – временное выражение, дата которого обозначена явно, понятна из контекста или общих знаний истории (Рисунок 42).

Для сущностей **Date: absolute** доступны следующие атрибуты:

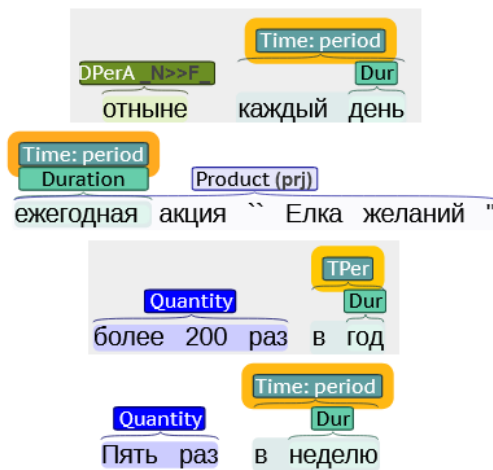


Рис. 41: Time: period – Периодичность



Рис. 42: Примеры разметки Date: absolute

- Past ($_P_$) – прошедшее время
- Present ($_N_$) – настоящее время
- Future ($_F_$) – будущее время

В зависимости от того, когда по отношению к текущему моменту⁶ происходят события, необходимо выбрать один нужный атрибут, если это время можно понять из контекста.

Date: relative – временное выражение, которое выражает когда, относительно текущего времени или другого события, происходит действие. Например, *впоследствии*, *ранее*, *позднее*, *через несколько часов*, *спустя два года*, *неделей раньше* и др.

Для сущностей **Date: relative** доступны те же атрибуты **Past**, **Present** и **Future**.

Date: relative может также обозначать время, похожее на [*после предъявленных обвинений*]_{Time: absolute}, однако в этом случае событие выражено неявно,

⁶Будем считать, что текущий момент – время написания статьи.

например: [после этого]_{Date:relative}, [после случившегося]_{Date:relative}, [после которого]_{Date:relative} и т. п.

Примечание. Одинокое упоминание времени конкретного дня нужно обернуть в `Date: relative` (Рисунок 43).

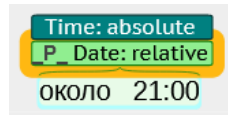


Рис. 43: Время в `Date: relative`

Date: period absolute – временной период, границы которого обозначены явно, понятны из контекста или общих знаний истории (Рисунок 44). Действие происходит на протяжении всего периода.

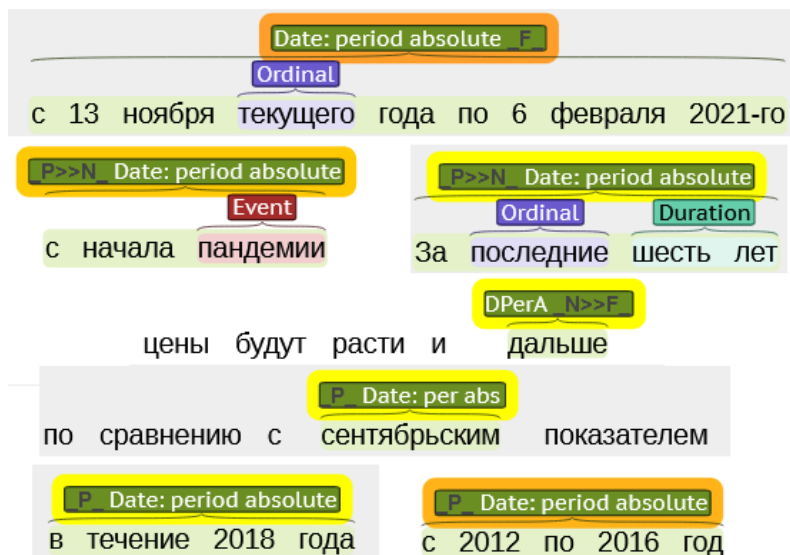


Рис. 44: Примеры разметки `Date: absolute`

Date: period absolute выделяется до ближайшего существительного слова и может пересекаться с другими сущностями.

Атрибуты для сущностей **Date: period absolute** обозначают когда начался и когда закончился период. Доступны следующие атрибуты:

- **Past** (`_P_`) – период начался и закончился в прошлом
- **Past-Pres** (`_P>>N_`) – период начался в прошлом, а закончился в текущий момент
- **Past-UNK** (`_P>>UNK_`) – период начался в прошлом, но неизвестно, когда он закончился/закончится
- **Past-Fut** (`_P>>F_`) – период начался в прошлом, а закончится в будущем
- **Pres-Fut** (`_N>>F_`) – период начинается в текущий момент, а закончится в будущем

- UNK-Fut ($_UNK \gg F_$) – неизвестно, когда период начался, но закончится он в будущем
- Future ($_F_$) – период начнется и закончится в будущем

Кроме того, есть два атрибута, которые указывают на продленность действия/события в прошлое (влево) или в будущее (вправо) от текущего момента.

- LCont ($>>>:$) – в тексте период указан от настоящего момента, но понятно, что действие/событие этого периода уже продолжается какое-то время.
- RCont ($:>>>$) – в тексте период указан до настоящего момента, но понятно, что действие/событие этого периода будет продолжаться еще какое-то время (Рисунок 45).

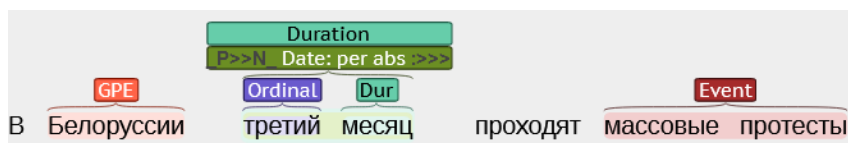


Рис. 45: Date: period absolute с атрибутом RCont

Date: period relative – реальный временной период, даты которого не выражены явно. (Рисунок 46). Действие происходит на протяжении всего периода.

Date: period relative выделяется до ближайшего существительного справа и может пересекаться с другими сущностями.

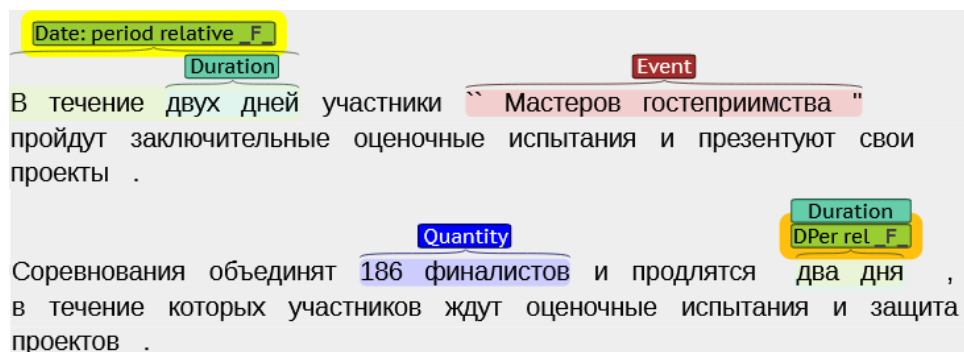


Рис. 46: Примеры разметки Date: period relative

Тип сущности **Date: period relative** имеет те же атрибуты, что и **Date: period absolute**. Время начала и конца определяется относительно текущего момента.

2.5.3 Date в Event и наоборот

Сущности **Event** могут включать в себя **Date** (Рисунок 47).

И наоборот, сущности **Date** могут включать в себя **Event** (Рисунок 48).



Рис. 47: Пример Event, содержащий Date



Рис. 48: Пример Date, содержащий Event

2.5.4 Ordinal в Date

Слова *первый*, *второй* и т. д., а также *текущий*, *последний*, *следующий*, *прошлый*, *нынешний*, *ближайший*, *настоящий*, *этот* и т. п. выделяются как **Ordinal**, если употребляется внутри **Date** (Рисунок 49).

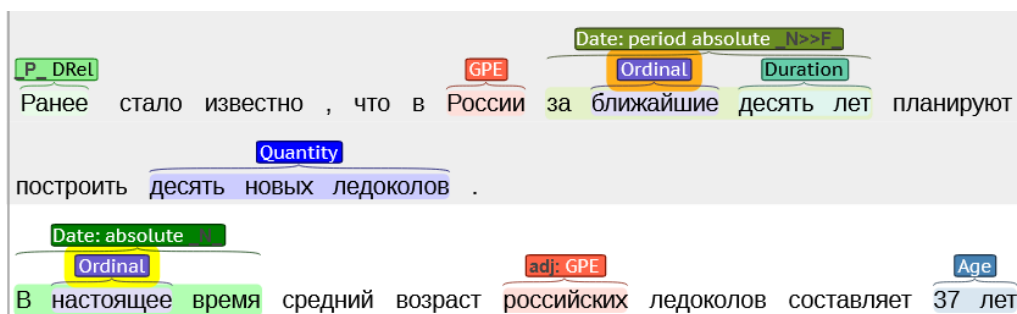


Рис. 49: Пример Date, содержащий Ordinal

2.6 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 территориальная граница и Правительство России, или в тексте есть уточнение, о столице какой страны идет речь (например, “американская/белорусская столица” или в абзаце указано название страны), “столица” размечается как **GPE**, “граница” – как **Facility**,

а “*правительство*” – как `Organization`.

Кроме того, одиночные сущности `Facility`, `Event` и `Service` могут быть выделены как именованная сущность, если в рамках одного абзаца с ними есть конкретизирующее уточнение (Подробнее в Разделе 2.1.1).

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

3 Разметка кореференции в brat

Все упоминания в `brat` отмечаются меткой `[X]`. На Рисунке 50 показаны выделенные упоминания в тексте.

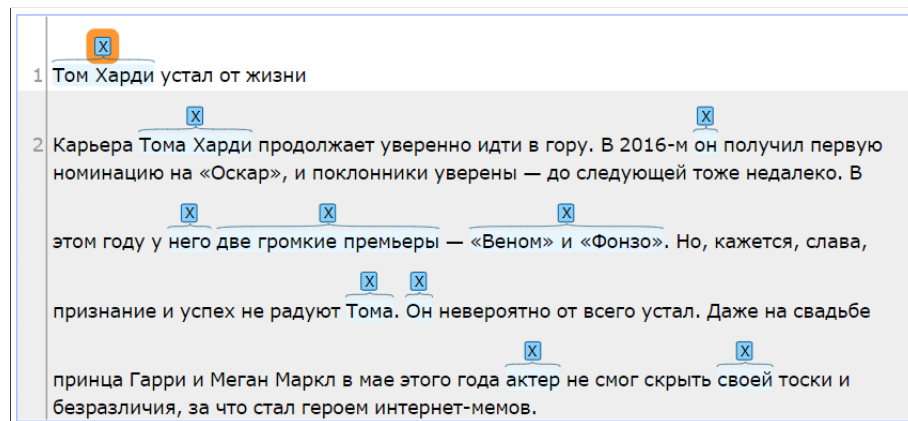


Рис. 50: Пример выделения упоминаний в `brat`

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

На Рисунке 51 показан пример текста с размеченными связями между упоминаниями.

Если в тексте нет референта, и все упоминания выражены местоимениями или нереферентными именными группами, следует отметить все упоминания такой кореферентной цепочки атрибутом **NoRef** (Рисунок 52).

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.

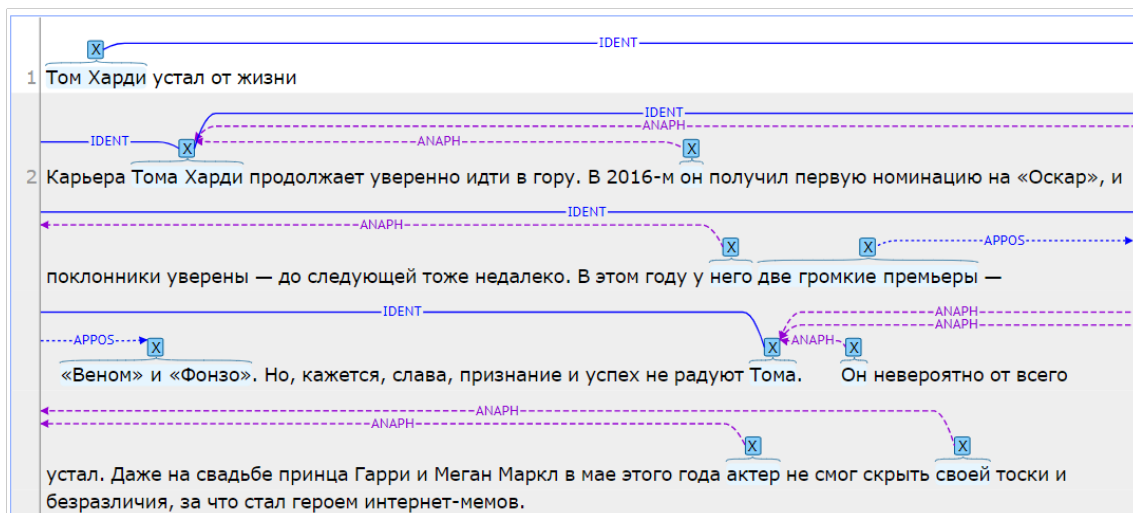


Рис. 51: Пример кореферентной разметки в brat

Text

Он

Search

Google, Wikipedia

Entity type

☒ X

☐ INVALID DOCUMENT !!!

Entity attributes

☒ NoRef

Рис. 52: Выбор атрибута **NoRef** для нереферентных упоминаний

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.

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

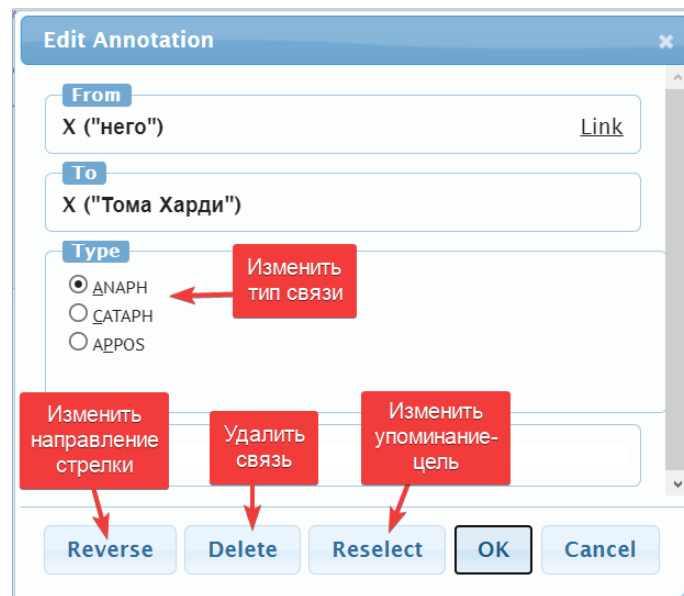


Рис. 53: Изменение или удаление связи между упоминаниями

4 Возможные ошибки и предупреждения

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

- **Пересекающиеся сущности.** Пересекаться с другими сущностями могут только сущности `Quantity` и `Date: period abs/rel`. В остальных случаях при пересечении метки подсвечиваются красным (Рисунок 54). Необходимо переразметить сущности таким образом, чтобы они не пересекались.

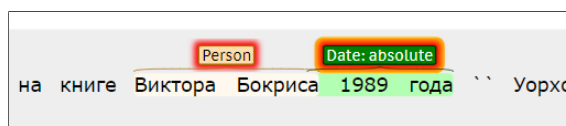


Рис. 54: Пример ошибки при пересечении сущностей.

Ошибки также возникают, если на пересечении вложенных сущностей оказываются лишние пробелы. Такие случаи тоже необходимо переразметить.

- **Пробелы в начале сущности.** Если в начало выделенной сущности попали пробелы, высветится предупреждающее сообщение (Рисунок 56). Чтобы предупреждение пропало, необходимо переразметить сущность (с помощью **Move** или **Delete**), при выделении не включая лишние пробелы в начало.

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

Person property Person
GPE
Госсекретарь Мичигана Джоселин Бенсон назвала иск штаба президента США Дональда Трампа

Organization
Person property Person
GPE

Text-bound annotation text ".
" shorter than marked span(s) [(788, 790), (791, 792), (793, 799)]

Unable to parse the following line(s):
3: T2 Person 788 790;791 792;793 799 .
4:
5: Бенсон

Рис. 55: Лишние пробелы на пересечении сущностей вызывают ошибки.

Time: absolute
5 Встреча длилась около 4 часов .

WARNING
The fragment [518, 537] (около 4 часов) is not contained in its designated chunk [520, 537] most likely due to the fragment starting or ending with a space, please verify the sanity of your data since we are unable to visualise this fragment correctly and will drop leading space characters

Рис. 56: Пример предупреждения при наличии пробелов в начале сущности.

- **Ошибка токенизации.** Могут встречаться тексты, в которых цельные предложения разбиты на несколько строк. Как размечать такие случаи, описано в разделе 1.5.

A Changelog

24-26.02.2021

1. Обновлена Таблица 1:

- В примеры **Person Property** добавлен “террорист”.
- Полностью переделаны разделы **Date** и **Time**, от них отделен **Duration**
- В **Nationality** добавлен атрибут **Resident** с описанием и примерами.
- Добавлена новая категория **Person Name**.
- В **Person** добавлены атрибуты **Male** и **Female**.
- Добавлены новые **Hotkey**.
- В **GPE** и некоторые подтипы **Social group** добавлен атрибут **Adjective**.
- В **Political Group** и **Religious group** добавлен атрибут **Source**.

28.02.2021

1. Полностью пересмотрена структура документации, составлено новое содержание по которому сегодня и в ближайшие дни будут вноситься изменения.
2. **Полностью переписан Раздел 1:** удалена неактуальная информация, изменена структура, добавлены новые детали, скриншоты. Часть информации перетянута из других разделов при реструктуризации.
 - Из описания раздела 1 удалена информация про необходимость использования браузера Mozilla Firefox.
 - Переделан раздел 1.3. Перетянут абзац про отображение текста в интерфейсе, подраздел про алиасы лейблов.
 - Перемещен и переписан раздел 1.3.2. Основная идея – можно выделять несколько строк сразу.
 - В раздел 1.3.4 добавлена информация про системные горячие клавиши **Delete** и **Insert**.
 - В раздел 1.5 дописана часть про выделение ошибок токенизации, добавлено два скриншота с примерами.
 - В раздел 1.7 добавлена информация про поиск во вкладке **Entity** со скриншотом. Добавлена сноска про три пробела.

28.02.2021

1. Полностью переписывается Раздел 2:

- Исправлено описание раздела 2.
- Изменены все скриншоты на Рисунке 19.

01.02.2021-02.02.2021

1. **Полностью переписан Раздел 2:** Пересмотрен и исправлен весь имеющийся текст, удалены неактуальные разделы, добавлены новые разделы и подразделы, вставлена актуальная Таблица 1.
 - Заменены все неактуальные скриншоты.
 - Добавлено более 20 новых скриншотов с примерами.
 - Удалены все неактуальные разделы.

- Добавлены разделы с пояснениями и примерами для наиболее частотных существей.
2. **Удален Раздел 4 (Отдельные случаи).** Вся необходимая информация перекочевала в другие разделы, неактуальная – удалена.
 3. **Обновлена Таблица 1.** Добавлены новые примеры.
 4. В разделе 1.7 добавлено примечание про то, как пользоваться функцией Concordancing и зачем она нужна.