

ru:corner

Разметка именованных сущностей и кореференции в brat

Анастасия Никифорова	Сергей Терновых
steysie@gmail.com	fostroll@gmail.com

Денис Киреев	Константин Ремизов
dkireev.71@gmail.com	mr.enslin@mail.ru

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
2.2	Вложенность или отдельные сущности?	14
2.3	Атрибуты типов сущностей	14
2.4	Общие правила разметки именованных сущностей	16
2.4.1	Уточняющие прилагательные	16
2.4.2	Уточняющие именные группы	16
2.4.3	Уточняющие предлоги и наречия	16
2.4.4	Знаки препинания	16
2.4.5	Кавычки	17
2.5	Особенности разметки некоторых типов сущностей	17
2.5.1	Person с вложенными Person Name, Person Property и др.	17
2.5.2	Person без имени	19
2.5.3	Family с Person и Person Property	19
2.5.4	GPE и Loc. Атрибут adj	23
2.5.5	Разметка Event	24
2.5.6	Выбор между Service, Service:Media, Org и Org:Media	25
2.5.7	Product: продукт, серия, уникальное название – как быть?	26
2.5.8	Numeric: Quantity. Пересечения Quantity	26
2.6	Разбор Time, Date и Duration	27
2.6.1	Подтипы Time	27
2.6.2	Подтипы Date	28
2.6.3	Date в Event и наоборот	32

2.6.4	Ordinal в Date	32
2.7	Одиночные именованные сущности и анафора	33
3	Разметка кореференции в brat	33
4	Разное	36
4.1	Особенности разных жанров	36
4.2	Возможные ошибки и предупреждения	37
4.3	Как быстро выделить текст без лишних пробелов	38
Приложение А Changelog		40
Приложение В Проверка аннотации поиском в brat		42

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).

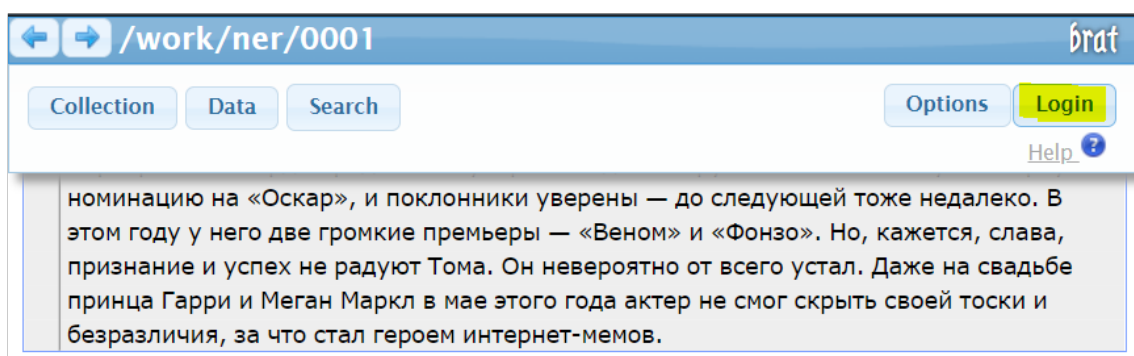


Рис. 1: Авторизация в brat

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

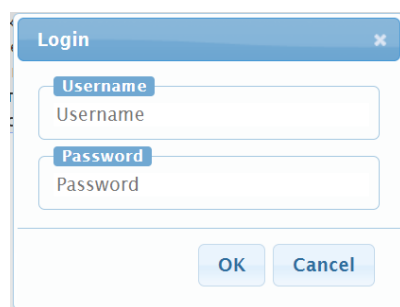


Рис. 2: Окно авторизации

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.

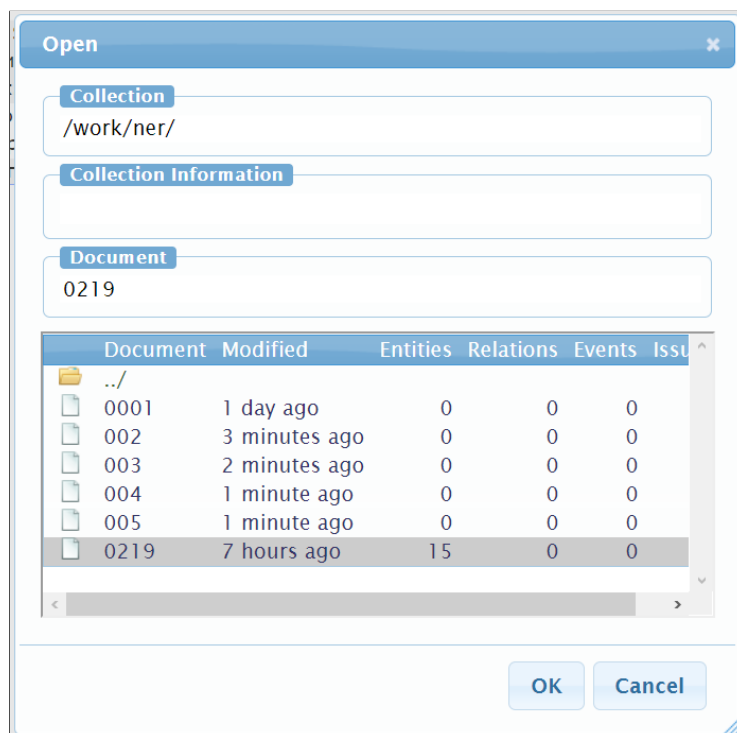


Рис. 3: Пример окна Collections

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.

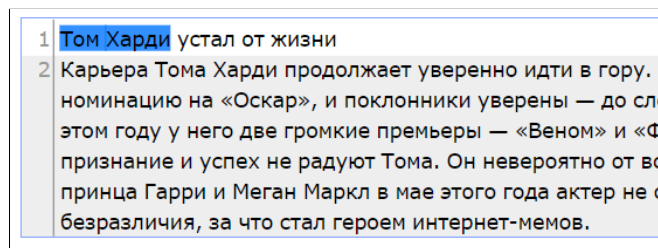


Рис. 5: Пример выделенного отрезка текста

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 \rightarrow или кликните на кнопку \Rightarrow 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.

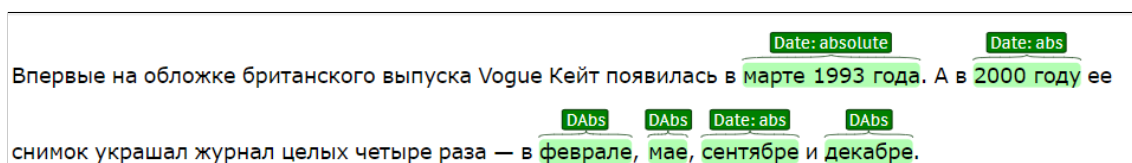


Рис. 7: Примеры алиасов сущности Date: absolute

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.

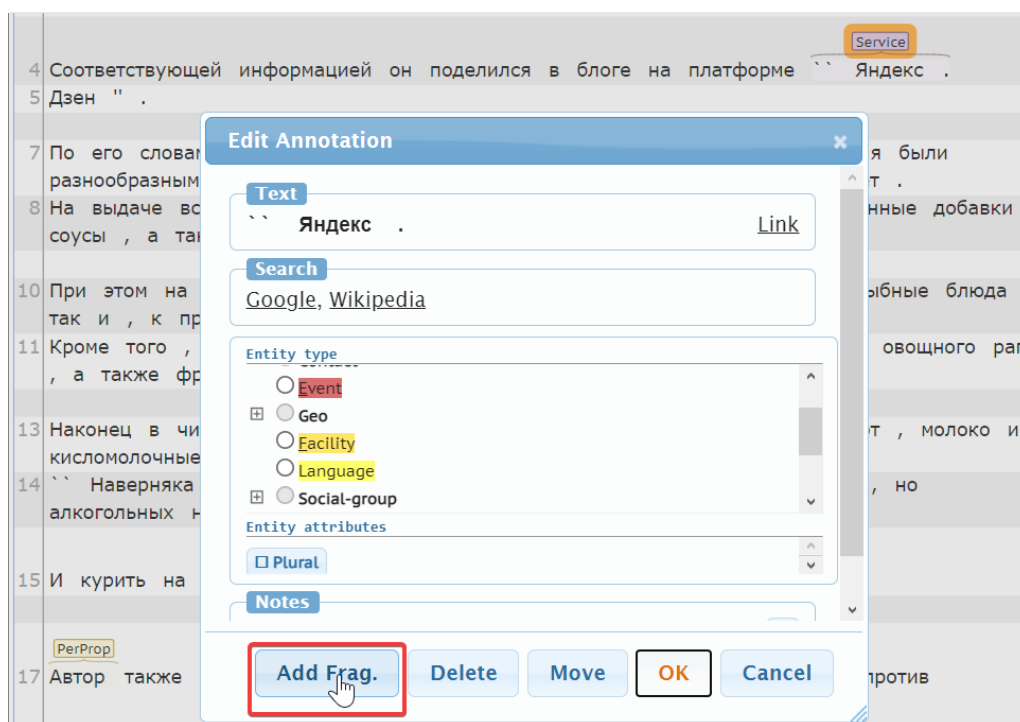


Рис. 8: Добавление фрагмента к сущности

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

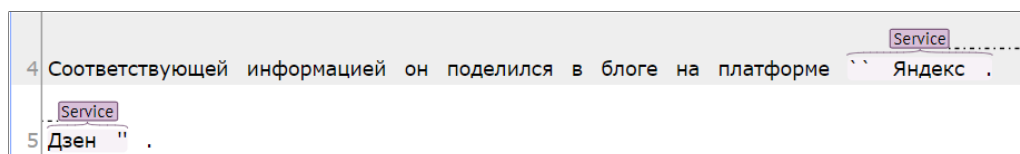


Рис. 9: Пример сущности из нескольких фрагментов

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).

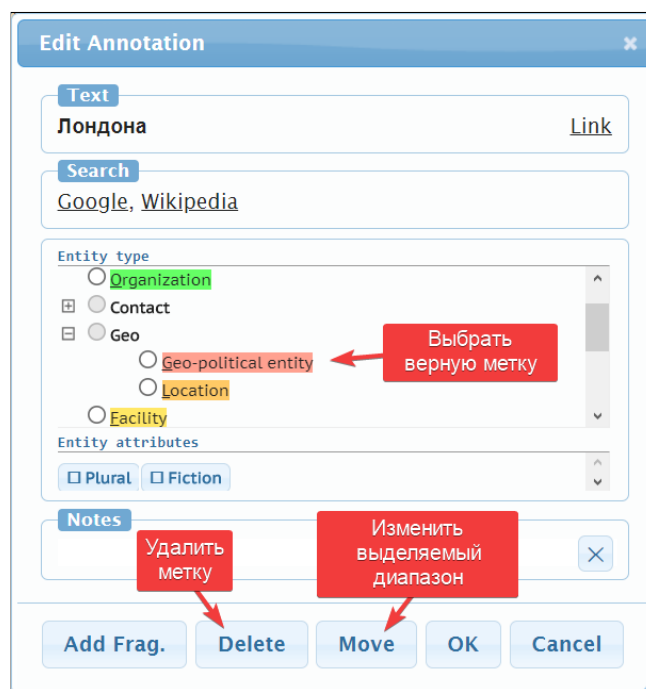


Рис. 10: Изменение и удаление меток именованных сущностей

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).

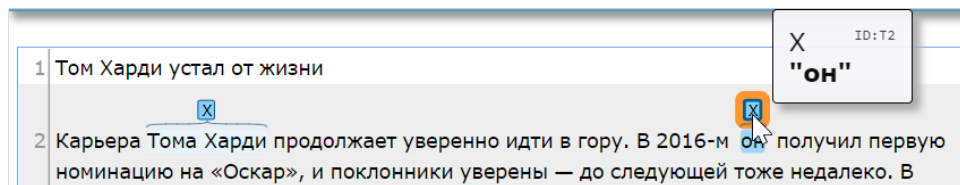


Рис. 11: Шаг 1. Выбор метки одного из упоминаний

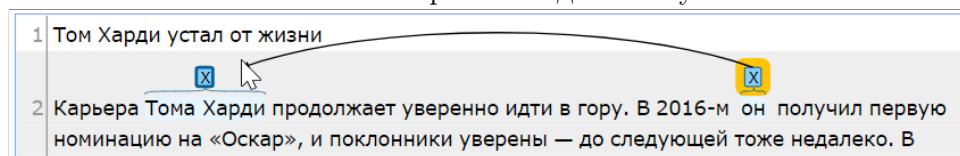


Рис. 12: Шаг 2. Связывание кореферентной группы

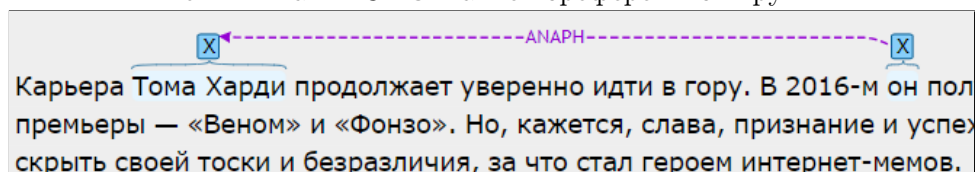


Рис. 13: Шаг 3. Выбор связи

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.

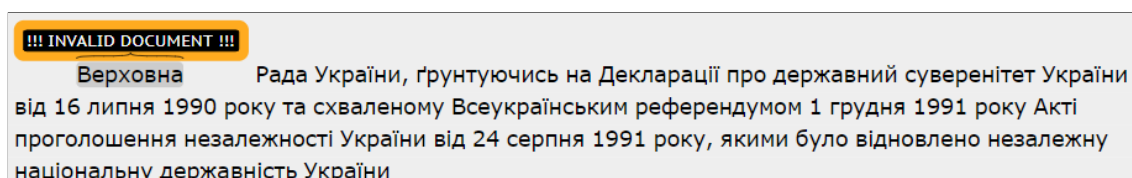


Рис. 14: Пример разметки нерелевантного текста

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).

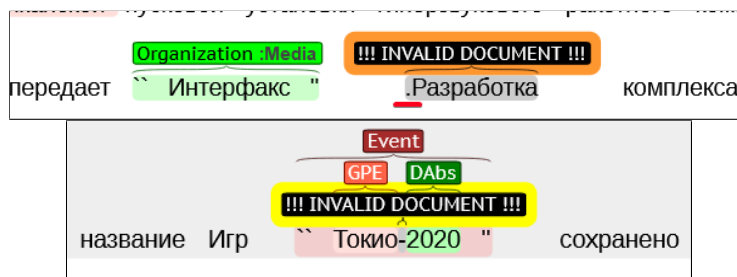


Рис. 15: Пример разметки ошибки токенизации

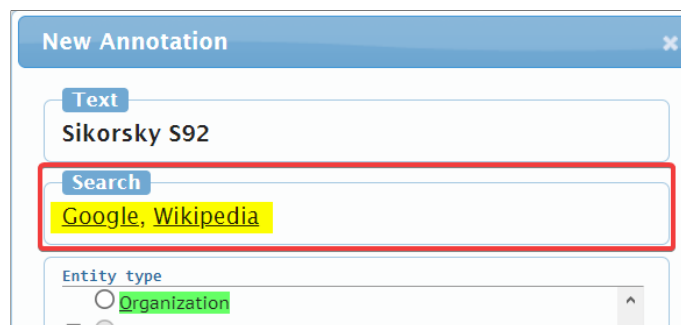


Рис. 16: Поиск незнакомых терминов в Google и Wikipedia

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.

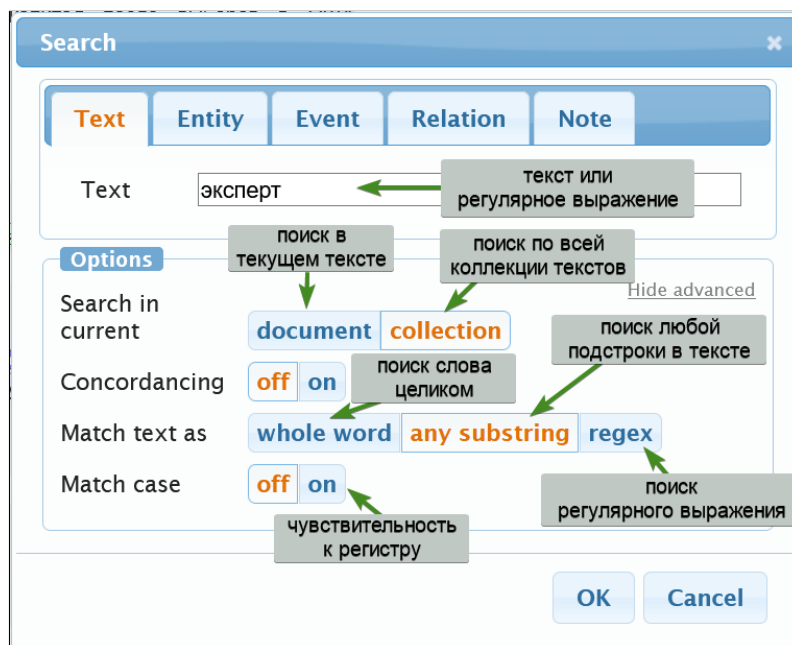


Рис. 17: Окно поиска. Вкладка Text

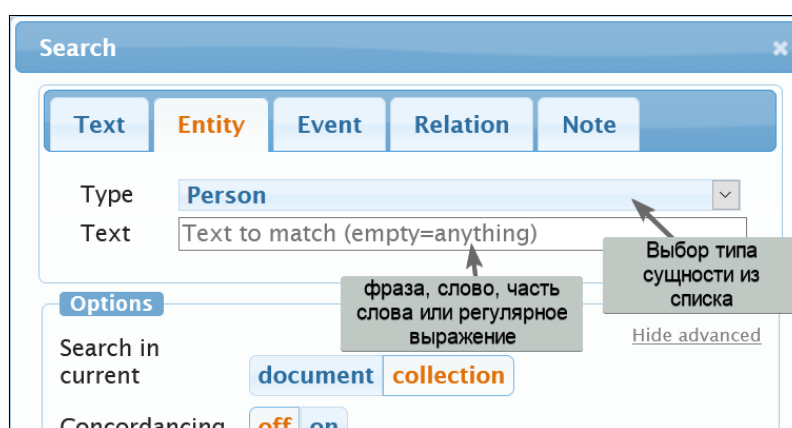


Рис. 18: Окно поиска. Вкладка Entity

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 (Рисунок 19). С помощью этой опции можно быстро найти нужную фразу в нужном документе, изучив окно с результатами поиска.

Document	Annotation	Type	Left context	Text	Right context
0002	T44	Family	стенах . Один , без	ребенка	и жены " , - рас
0002	T45	Family	дин , без ребенка и	жены	" , - рассказал он
0003	T35	Family	в Орловской области у	бабушки	и дедушки , которые
0003	T36	Family	области у бабушки и	дедушки	, которые воспитывали
0003	T37	Family	али его после развода	родителей	. Злоумышленника задержал

Рис. 19: Окно поиска с опцией **Concordancing: on**. Пример с поиском сущностей **Family**

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.), а также ошибки токенизации или код страницы. Сокращения типа "т.к.", "г. Москва", "и. о. мэра" и др. в эту группу не относятся	Отметьте первое слово в нерелевантном тексте. В случае ошибки токенизации или кода страницы, выделите эти участки
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, Андрияха, 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-чье-имя-нельзя-называть
		<input checked="" type="checkbox"/> Unconscious		Любые имена, которыми люди называют (персонифицируют) неразумные уникальные объекты: животные, техника, игрушки и т. п.	Барсик, Белка, Стрелка, Несси (лох-несское чудовище) + имена, которые люди дают своей технике, машинам, игрушкам и т. д.
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, сэр, господин, ее высочество, леди, сквайр, II, старший/младший (напр., Трамп 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, обвиняемый, задержанные, подозреваемый, пенсионер, убийца, активист
		<input checked="" type="checkbox"/> Plural		Несколько Property в одной сущности	"главы России, Азербайджана 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, армия, власти, руководство, сборная
		<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, Спектр, 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, Алтуньевский 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, столичный, отечественный
	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
	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, гречанка, 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
Social-group	Family		Y	Обозначения родственных связей (слово "брак" не размечаем)	мать, брат, старший сын, сводная сестра, двоюродная тетя, брат жены, семья, род, сожители, пара, молодожены, опекун, приемный отец
	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", тролль

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, по-русски, по-английски
		<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 <input type="checkbox"/> Unique	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 Эйвон 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
		<input checked="" type="checkbox"/> Unique		Уникальные объекты, которые имеют свои названия: корабли и суда, космические корабли и др.	Броненосец "Потемкин", эсминец Джон Маккейн, корабль "Сметливый", баркас "Fishizzle", СОЮЗ МС-16, ледокол "Арктика"
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, сайт, твиттер-аккаунт "Котики", в т. ч. одиночные "пост", "страница", "паблик" 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				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
Date (раздел 2.6.2)	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 Ordinal "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, после концерта, на 7-м месяце беременности, скоро, недавно
	Date: relative		Z	Время относительно текущего времени или другого события, когда происходит событие/действие	после случившегося, впоследствии, до этого, после которого, ранее, позднее, через несколько часов, спустя два года, на 59-й минуте (не указано, о каком событии идет речь)
	Date: period absolute	обязат-но 1 из <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 <input type="checkbox"/> Fut <input type="checkbox"/> LCont <input type="checkbox"/> RCont	X	Реальный период на временной шкале (атрибуты указывают на начало и конец периода, см. раздел 2.6.2)	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", с самого начала матча (событие указано явно)
	Date: period relative		C	Реальный временной период, даты которого не выражены явно (в этот период что-то реально происходит, не абстрактный период)	в течение двух дней, чуть более года, два часа подряд, с самого начала (не указано о "начале" какого события идет речь)

Таблица 1. Типы сущностей, их описания и примеры

Категория	Тип сущности	Атрибуты	Hotkey	Описание типа сущности	Пример
Time (раздел 2.6.1)	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 (одиночное; если не указано, после чего)
	Time: period		B	Период времени, абстрактный или гипотетический (общие факты, предположения), а также периодичность	(30 дней отпуска) в год, за два года (можно выучить китайский), (рана заживает) спустя 10-20 дней, (чиним ноутбук) за четыре дня, ежедневный, ежемесячный, ежегодный, каждый день
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	Порядковые числительные и слова, указывающие очередность объектов (какой по счету?)	первый, второй, тысячный, следующий, предыдущий, (пред)последний, повторно, первичный, впервые, не раз, еще раз
	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, 1987 года рождения
	Quantity		8	Измерения, количество чего-либо с единицей измерения (сколько?), включая счета матчей и фразы с "несколько" и "целый ряд"	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, втройне, однажды (в значении "один раз"), пара кексов, тройка лошадей, единственный, топ-10, 5:2 (счет матчей), несколько, еще раз, трехкратный, целый ряд текстов, дуэт, трио, квартет
Other term		<input type="checkbox"/> Plural <input type="checkbox"/> Fiction	Q	Термины, не подпадающие под категории выше, включая награды, 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 `gui:corner` 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 2.3.

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 20.

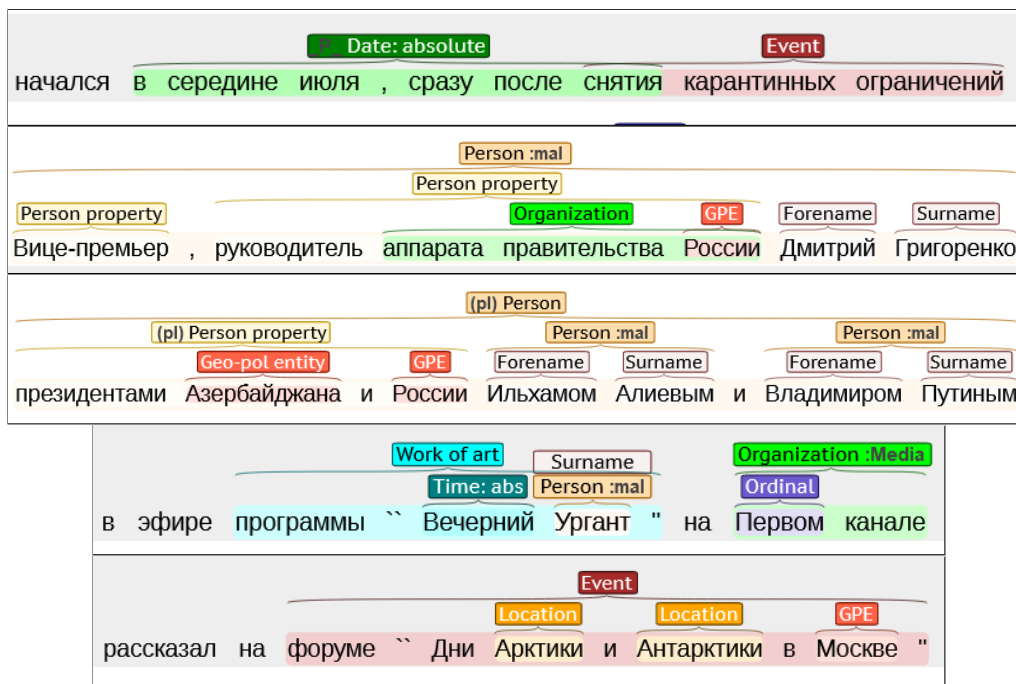


Рис. 20: Примеры пересекающихся и вложенных сущностей

В сущностях с атрибутом `Plural` по возможности размечаются отдельные вложенные сущности.

Примечание. Пересекающимися могут быть только сущности из категории **Quantity** (количество) и **Date**. Пример пересечения сущностей: “{шестерых [членов] экипажа}”. Можно ориентироваться на то, что пересечение неразрешенных типов сущностей подсвечивается красным. В таком случае, нужно использовать вложенность.

2.2 Вложенность или раздельные 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 раздельно, без вложенности (см. “Раздельные сущности” в Таблице 2).

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

Таблица 2: Примеры цельных, вложенных и раздельных 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”]

Подробнее о выделении одиночных сущностей **Organization**, **Facility**, **GPE**, **Location** и **Event** в Разделе 2.7.

2.3 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*),

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]]”

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, персонифицирующие неразумные объекты: животных, технику, автомобили, игрушки и т. п. Доступно только для **Person**.
- **Unique** – уникальные объекты, которые имеют свои названия: корабли и суда, космические корабли и др. Доступно только для **Product**.

По умолчанию ни один атрибут не выбран. Чтобы выбрать атрибут, после выбора типа сущности нажмите на один из атрибутов в поле **Entity attributes**. В соответствующем поле появится флажок, и атрибут подсветится оранжевым (Рисунок 21).



Entity attributes				
<input type="checkbox"/> Plural	<input type="checkbox"/> Fiction	<input type="checkbox"/> Unconscious	<input checked="" type="checkbox"/> mal	<input type="checkbox"/> fem

Рис. 21: Выбор атрибутов

Сущности групп **Date** и **Time** имеют свои атрибуты, которые подробно рассмотрены в Разделе 2.6.

2.4 Общие правила разметки именованных сущностей

2.4.1 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.

2.4.2 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.

2.4.3 Уточняющие предлоги и наречия

Уточняющие предлоги и наречия могут входить в ряд типов сущностей:

- Все типы сущностей групп **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 **Quantity** могут входить уточняющие предлоги, например: “[в трёх километрах от]_{Quantity} деревни”.

2.4.4 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 22).

Важно! В сущностях **Product** (напр., *Шанель*) и **Event** (*Ураган Катрина*) **Person Name** не выделяется.

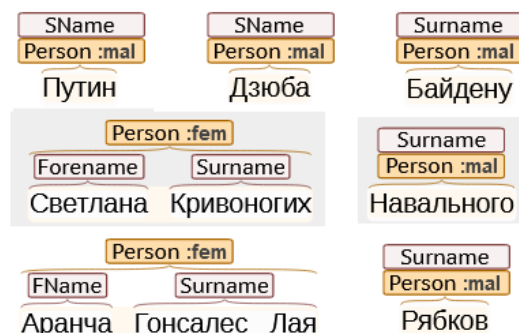


Рис. 25: Разметка одиночных имен: **Person+Person Name**

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 (Рисунок 26).

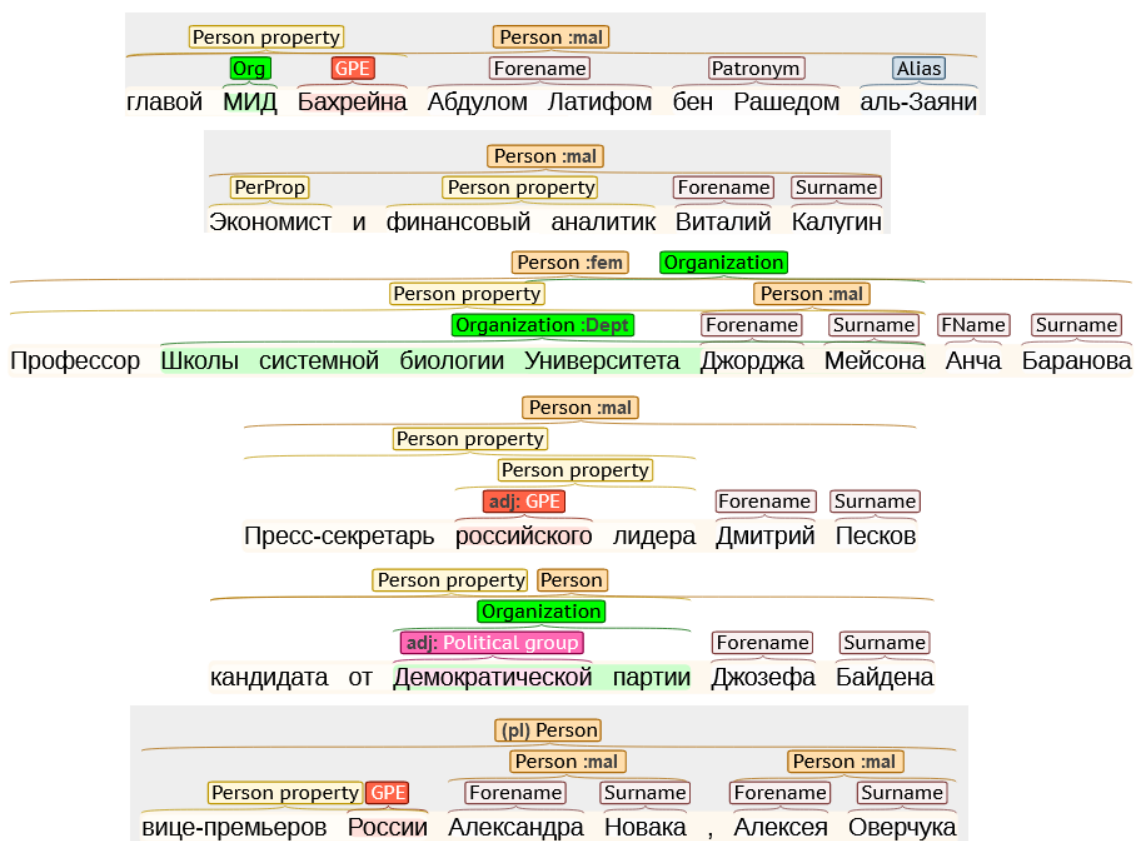


Рис. 26: Разметка вложенных **Person**

Исключением является сущность **Age**. Она не входит в сущность **Person**, если стоит перед ней (Рисунок 27). Если возраст указан в середине **Person**, он автоматически входит в **Person** (Рисунок 28).

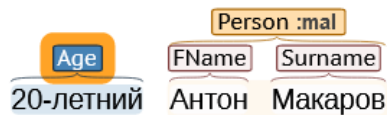


Рис. 27: Age перед Person

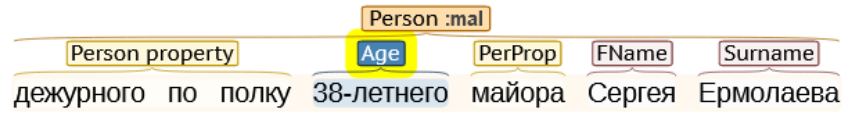


Рис. 28: Age внутри Person

Примечание. В основе **Person** может быть сущность **Family** (Раздел 2.5.3) или упоминания действующих лиц без имени (Раздел 2.5.2).

2.5.2 Person без имени

Person может быть выделен без имени, если обозначает конкретное действующее лицо или группу лиц, которое выполняет действие или над которым выполняется действие. Подразумевается, что у этих действующих лиц есть имена. Если возможно, уточняется атрибут пола **fem/mal**.

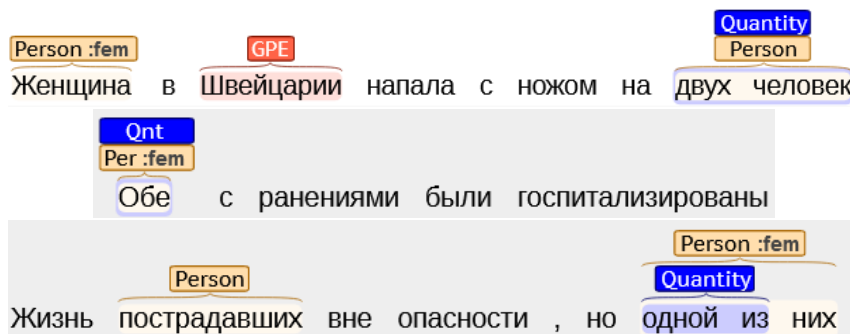


Рис. 29: Person без имени без вложенных сущностей

В отдельный **Person** могут выделяться отдельные слова или словосочетания (“женщина, мужчина, один из них и др.”, Рисунок 29) или сущности **Person Property**, **Family**, **Political group**, **Religious group**, **Nationality** и **Quantity** (Рисунок 30).

Слишком абстрактные упоминания группы лиц не выделяется как **Person**.

2.5.3 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.

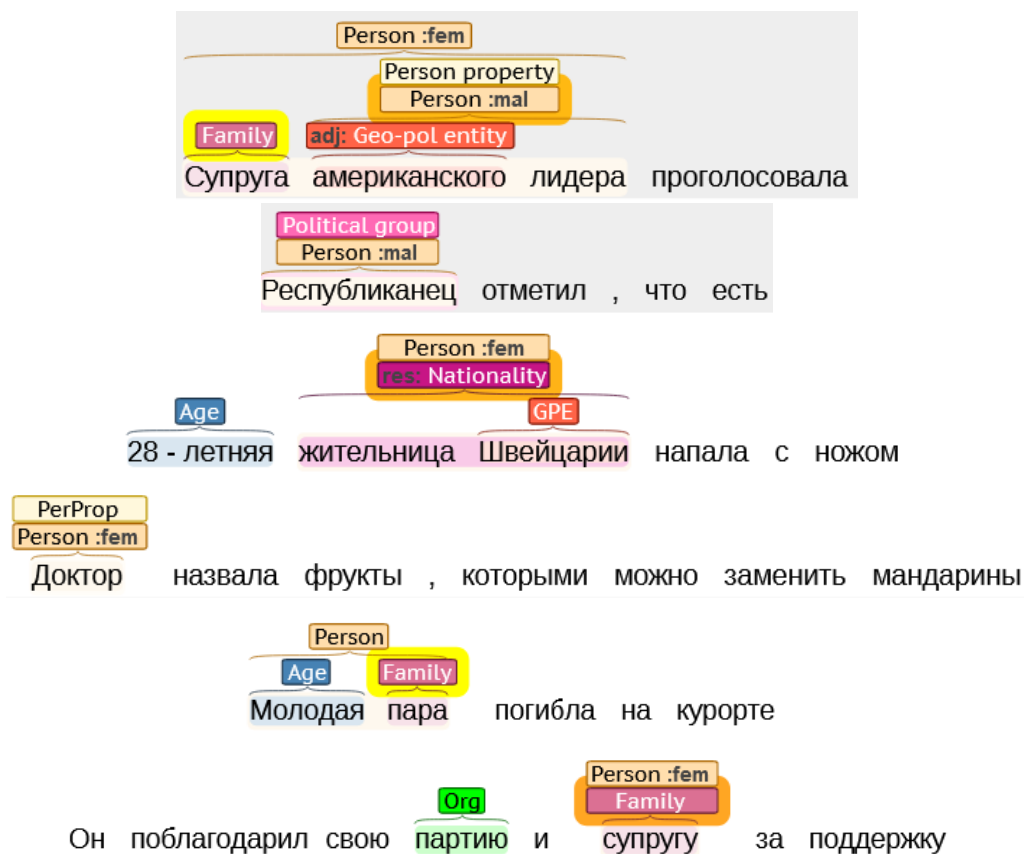


Рис. 30: Person без имени с вложенными сущностями

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

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

местоимение, оно включается в общую сущность **Person**. На Рисунке 31 приведены примеры разметки таких сочетаний.

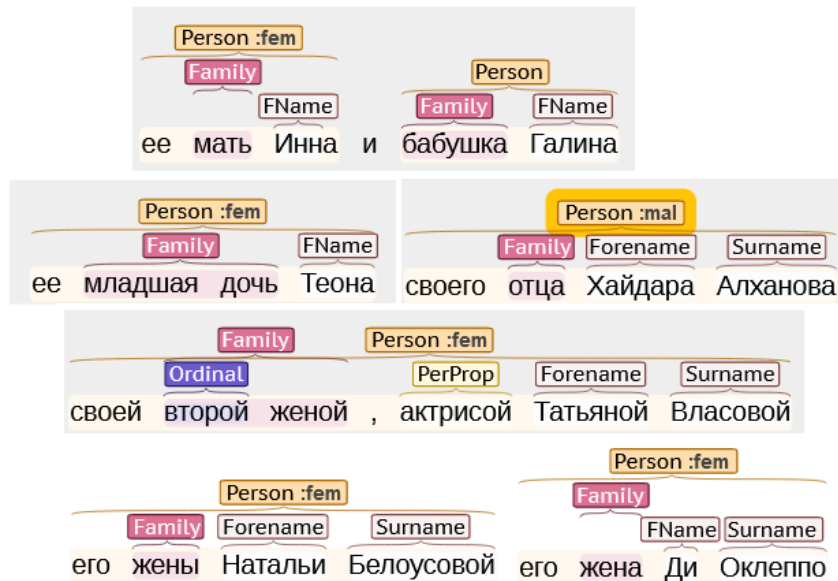


Рис. 31: Family + Имя в именительном падеже

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 (Рисунок 32). Если перед **Family** есть притяжательное местоимение, оно включается во внешнюю сущность **Person**.

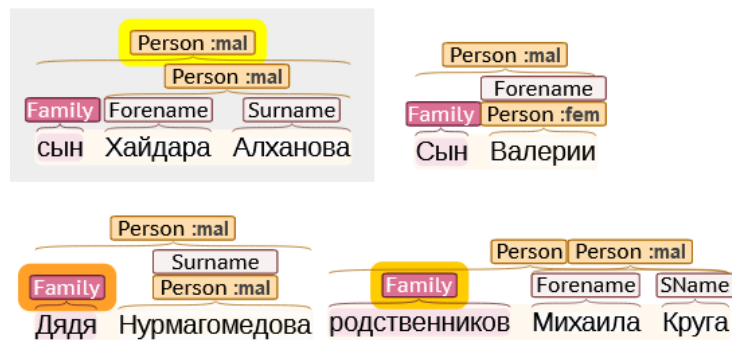


Рис. 32: Family + Имя в родительном падеже

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 33 приведены примеры разметки таких сочетаний.

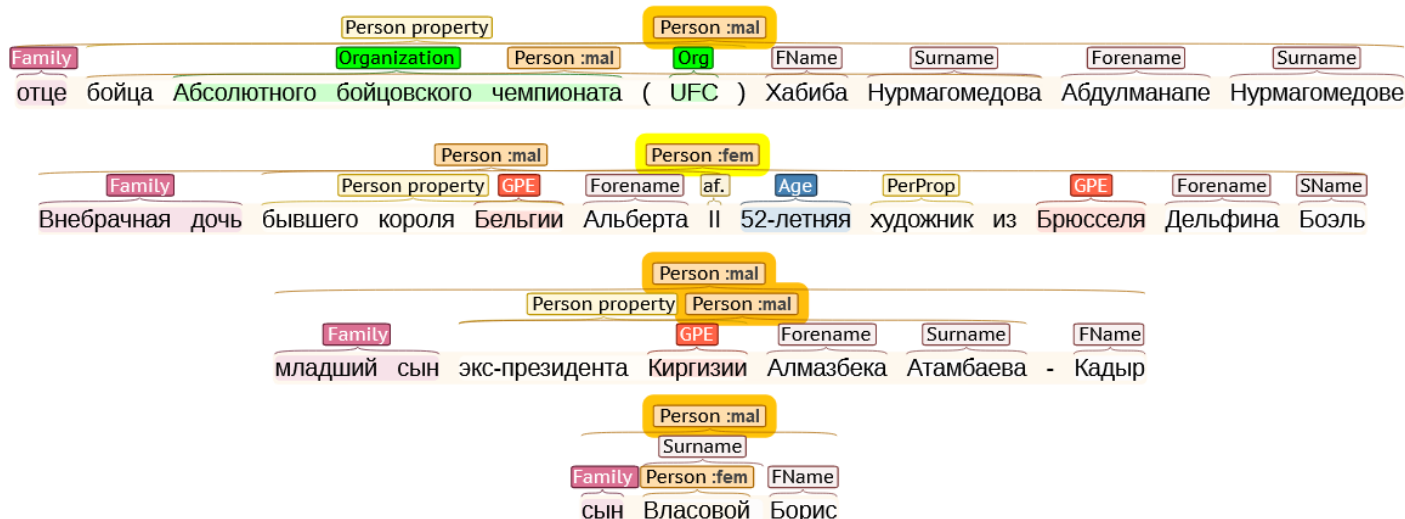


Рис. 33: Family + Имя в род. падеже + Имя в имен. падеже

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 34 приведены примеры разметки таких сочетаний.

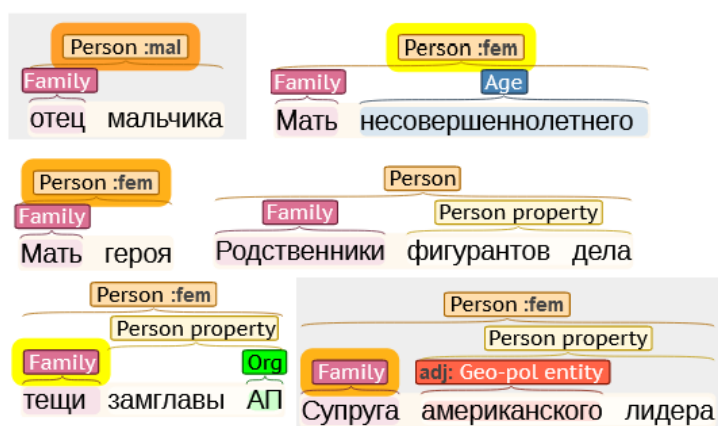


Рис. 34: Family + Property или похожее на Property слово в род. падеже

Притяжательное местоимение + Family

Пример: “*его брат*”. Такое сочетание объединяется в одну сущность **Person** (Рисунок 35). По возможности уточняется атрибут пола **fem/mal**. Исключениями являются слова, обозначающее цельное объединение родственников: *семья*, *род*, *династия* и т. п.

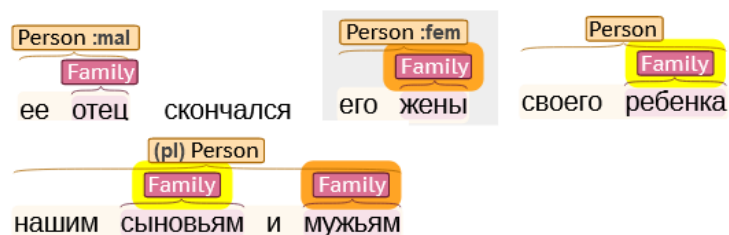


Рис. 35: Притяжательное местоимене + Family

Family и приложения/аппозитивные конструкции

Пример: “Иван, брат Бориса, ушел.”. Если сущность Family с зависимыми словами (в род. падеже) встретила в составе приложения или в качестве сказуемого в конструкциях типа “Иван – (это) сын Бориса”, она объединяется в отдельную сущность Person с зависимыми словами (Рисунок 36).

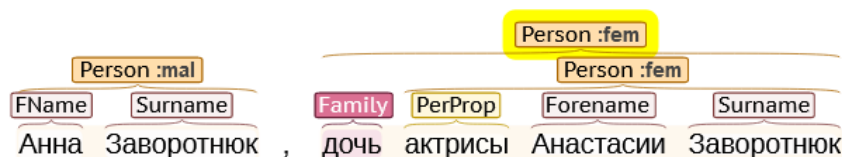


Рис. 36: Family и приложения/аппозитивные конструкции

Исключения

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

2.5.4 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 37).

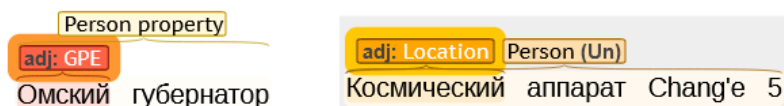


Рис. 37: Атрибут Adjective в GPE и Loc

Если имеется в виду сама местность, атрибут Adjective не ставится, даже если сущность выражена только прилагательным: либо уточняющее существи-



Рис. 38: Случаи, когда атрибут Adjective не ставится

тельное типа *область, республика, село* может быть опущено, либо это составные части Plural-сущностей (Рисунок 38).

Кроме того, как описано в Разделе 2.7, выделяются одиночные GPE-анафоры.

2.5.5 Разметка Event

Event включают в себя события самого разного рода: от мероприятий с конкретным названием (например, *чемпионат мира по футболу, XI Всероссийский конгресс пациентов*) до событий и процессов, выраженных фразой или одним словом (например, *нападение, запрет, санкции, поставка продукции, интервью, старт массовой вакцинации, рост случаев заболевания* и мн. др.).

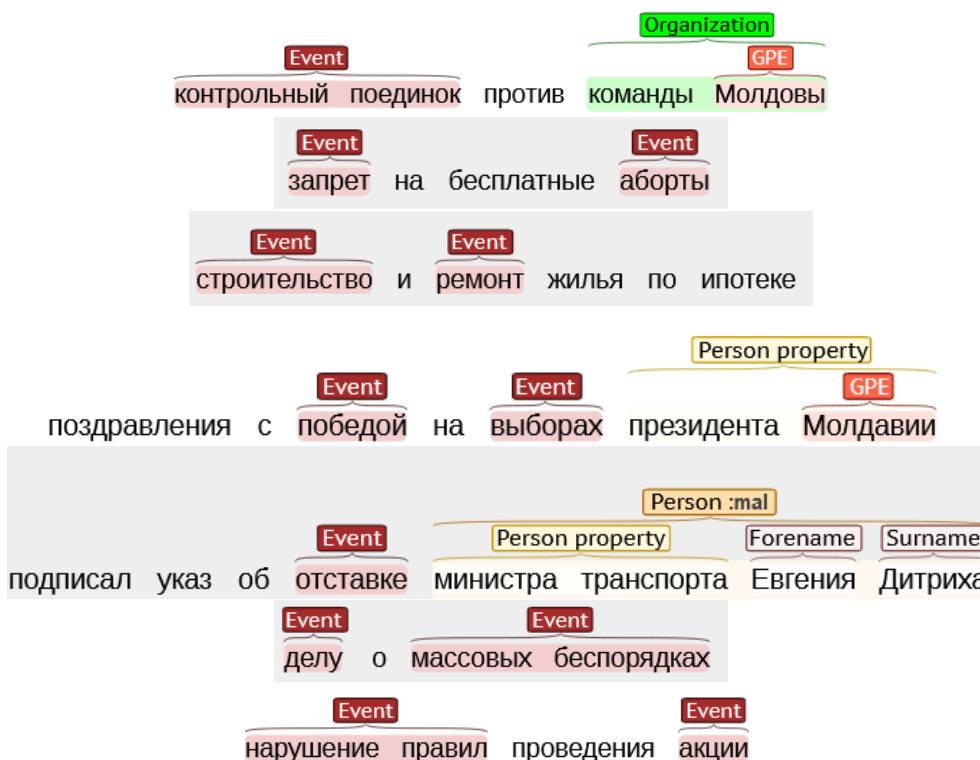


Рис. 39: Примеры выделения Event без названий

Сущности **Event**, как правило, являются конкретными и могут быть отнесены к какой-либо временной точке или промежутку времени, однако могут обозначать и абстрактные события без привязки к конкретному времени.

В случае событий без названия, выделяется основное слово-существительное с предшествующими прилагательными, а также, если таковые есть, уточнения вплоть до первого существительного после. Однако если первое существительное после является частью другого **Event**, этот **Event** захватывается целиком. В свою очередь, если первое существительное после является частью **Person**, **Person Property**, **Organization** или другой длинной сущности, выделяется только основное слово-существительное с предшествующими прилагательными.

Примеры выделения **Event** без названий изображены на Рисунке 39.

2.5.6 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.

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]

2.5.7 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.

Примечание. **Person:Unconscious** – уникальное название неодушевленных объектов или неразумных животных. 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.

2.5.8 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 40 выделенное выражение можно восстановить до “умер 7071 человек”.

Сущности **Quantity** захватывают конкретизаторы (подробнее в Разделе 2.4.3), численное количество и ближайшую зависимую именную группу, обозначающую считаемые объекты. **Quantity** могут пересекаться с другими сущностями (Рисунок 41).

зарегистрировано	450	436	новых заражений	, выздоровел	329	201	человек	умер	7071

Рис. 40: Пример единичного числа в роли Quantity

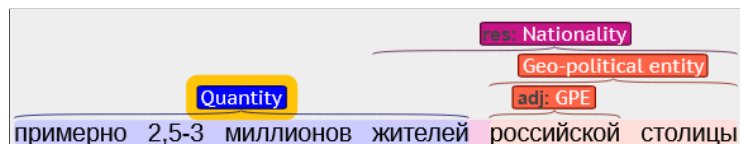


Рис. 41: Пример пересекающейся сущности Quantity

2.6 Разбор Time, Date и Duration

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

Time – это либо время внутри дня (*23:00, утро, темное время суток*), либо указание на конкретное время внутри некоего абстрактного периода.

Date – это конкретное положение (точка или период) на шкале времени.

В качестве **Date** могут упоминаться не только конкретные даты, но и указания на время относительно **событий**, в том числе включающие указатели “*этот, данный*” и т.п., которые следует выделить как **Ordinal**.

Например: после данного заявления, перед этим взрывом.

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 42).

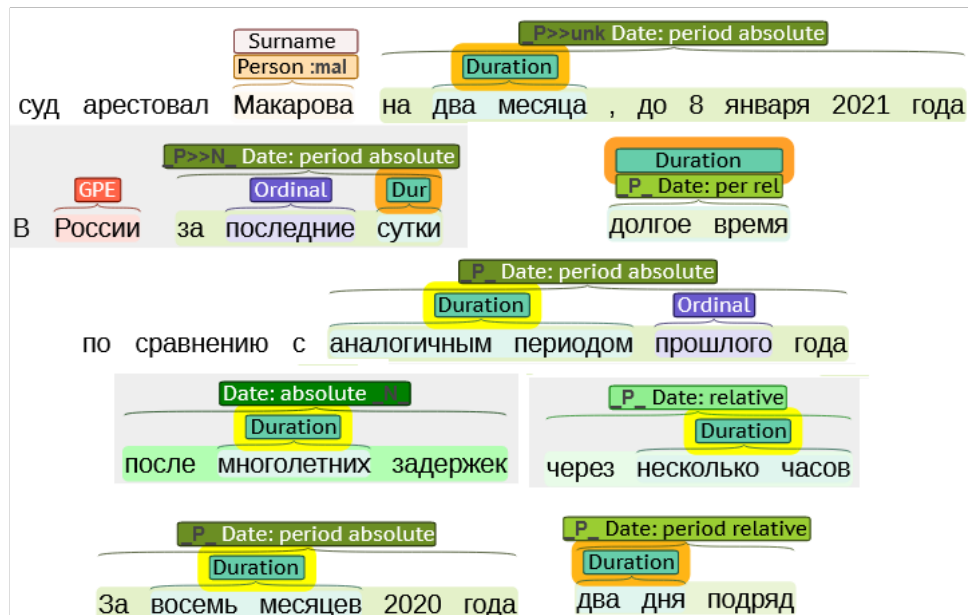


Рис. 42: Примеры разметки Duration

2.6.1 Подтипы Time

Time: absolute – время или период, когда что-то произошло/происходит/произойдет или может произойти конечное количество раз (не длится на про-

тяжении всего периода). Отвечает на вопросы *когда? во сколько?* (Рисунок 43).

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 44) или периодичность (Рисунок 45).

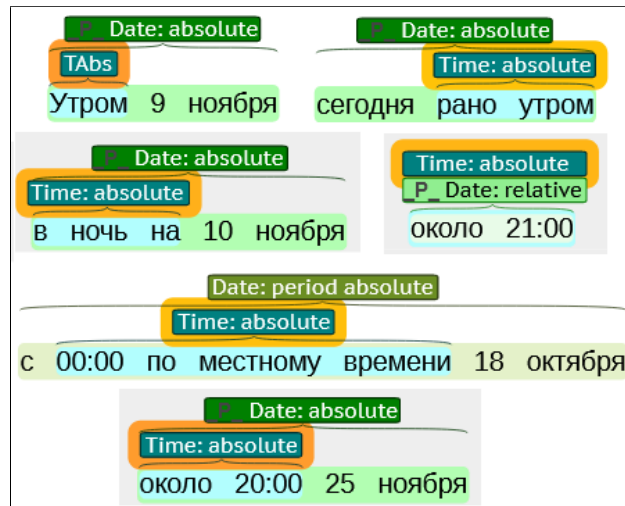


Рис. 43: Примеры разметки Time: absolute

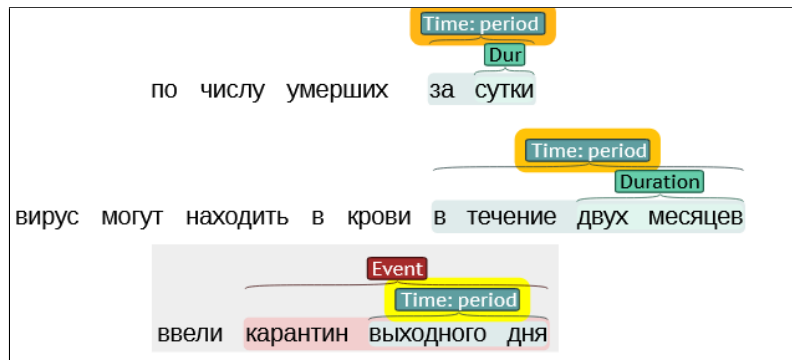


Рис. 44: Time: period – абстрактный период времени

2.6.2 Подтипы Date

Date: absolute – временное выражение, дата которого обозначена явно, понятна из контекста или общих знаний истории (Рисунок 46).

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

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

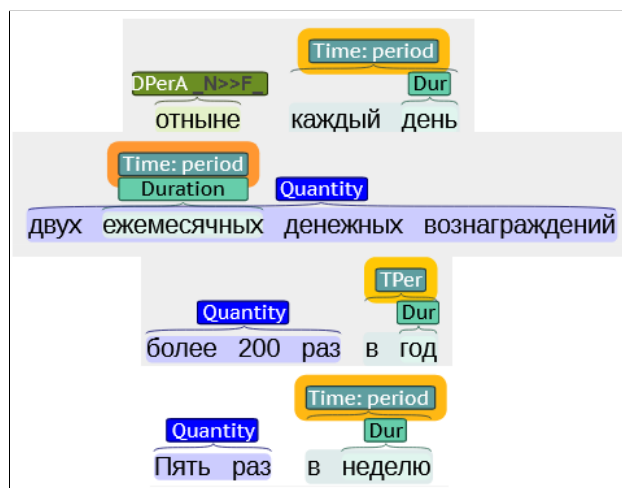


Рис. 45: Time: period – Периодичность



Рис. 46: Примеры разметки Date: absolute

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.

Примеры **Date: relative**: *впоследствии, ранее, позднее, через несколько часов, спустя два года, недель раньше* и др.

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

Date: relative может также обозначать время, похожее на [*после предъяв-*

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

ленных обвинений] $_{Time: absolute}$, однако в этом случае событие выражено неявно, например: [после этого] $_{Date: relative}$, [после случившегося] $_{Date: relative}$, [после которого] $_{Date: relative}$ и т. п.

Примечание. Одиночное упоминание времени конкретного дня нужно обернуть в `Date: relative` (Рисунок 47).

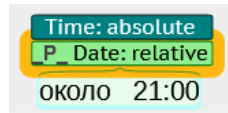


Рис. 47: Время в `Date: relative`

Date: period absolute – временной период, границы которого обозначены явно, понятны из контекста или общих знаний истории (Рисунок 48). Действие происходит на протяжении всего периода.

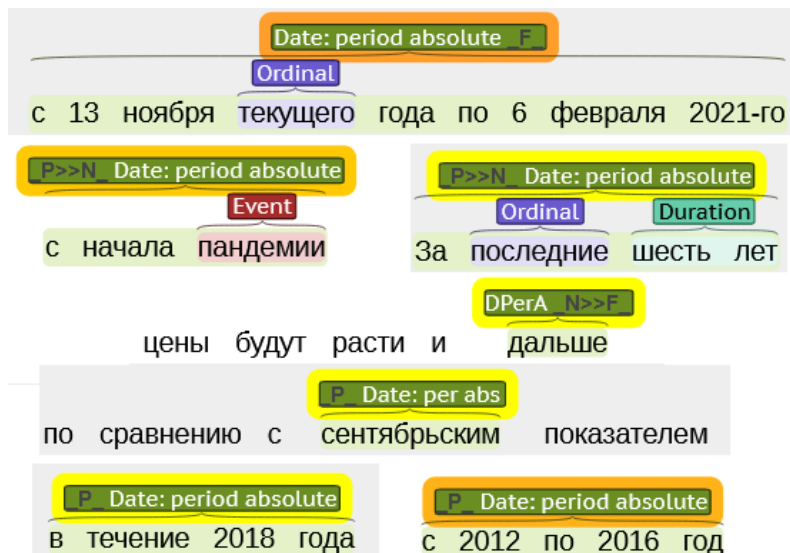


Рис. 48: Примеры разметки `Date: period absolute`

Date: period absolute выделяется до ближайшего существительного справа 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** ($_P\gg N_$) – период начался в прошлом, а закончился в текущий момент.
- **Past-UNK** ($_P\gg UNK_$) – период начался в прошлом, но неизвестно, когда он закончился/закончится. Например, “Ему назначили [20 лет] лишения свободы” – понятно, что период начался в прошлом, но неизвестно, закончился он или ещё нет.

- Past-Fut ($_P \gg F_$) – период начался в прошлом, а закончится в будущем.
- Pres-Fut ($_N \gg F_$) – период начинается в текущий момент, а закончится в будущем.
- UNK-Fut ($_UNK \gg F_$) – неизвестно начало периода, но его конец в будущем. Например, “*Полет продлится [два года]*” – ясно, что полет рассчитан на два года, но не ясно, он уже начался или начнется в будущем.
- Future ($_F_$) – период начнется и закончится в будущем.

Кроме того, есть два атрибута, которые указывают на продленность⁷ 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. Выбирается в дополнение к атрибуту $_Pres \gg Fut_$.

Например, “*Полет продлится [ещё два года]* $_Pres \gg Fut_ + LCont$ ” – ясно, что полет закончится через два года от текущей даты, но также ясно, что он начался в прошлом и 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. Выбирается в дополнение к атрибуту $_Past \gg Pres_$.

Например, “*Полет длится [уже два года]* $_Past \gg Pres_ + RCont$ ” – ясно, что начало полета началось в прошлом и продолжается до текущего момента, но также ясно, что он еще не закончился и продолжится в будущем. Ещё один аналогичный пример на Рисунке 49.

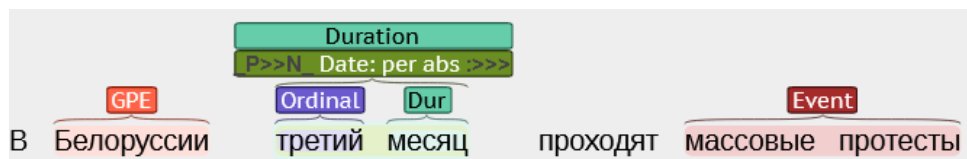


Рис. 49: Date: period absolute с атрибутом RCont

Date: period relative – реальный временной период, даты которого не выражены явно. (Рисунок 50). Действие происходит на протяжении всего периода.

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

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

⁷от англ. *continue* – продолжаться

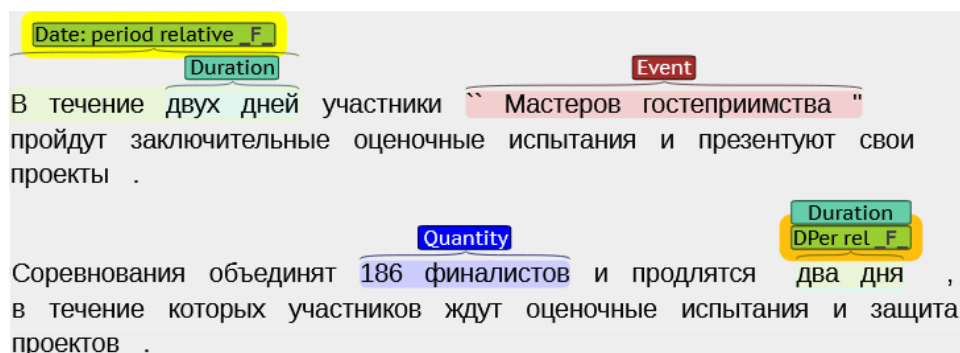


Рис. 50: Примеры разметки Date: period relative

2.6.3 Date в Event и наоборот

Сущности Event могут включать в себя Date (Рисунок 51).

И наоборот, сущности Date могут включать в себя Event (Рисунок 51).

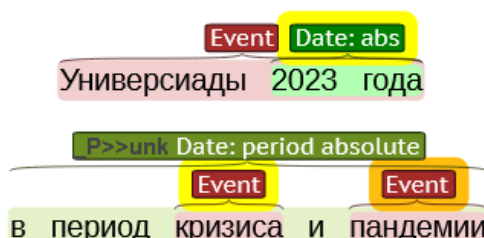


Рис. 51: Пример Event, содержащий Date, и наоборот

2.6.4 Ordinal в Date

Слова *первый*, *второй*, *последний*, *следующий* и т. д., а также *текущий*, *прошлый*, *нынешний*, *ближайший*, *настоящий*, *этот*⁸ и т. п. выделяются как Ordinal, если употребляется внутри Date (Рисунок 52).

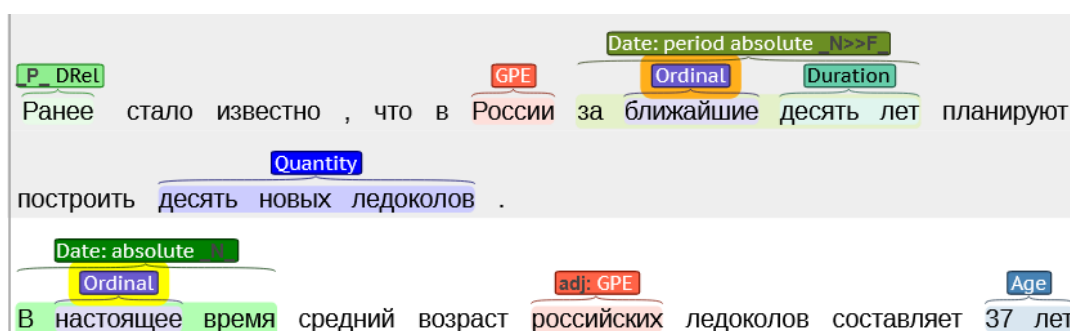


Рис. 52: Пример Date, содержащий Ordinal

⁸Эти слова выделяются как Ordinal только внутри дат. Вне дат они могут выделяться как Date: absolute, если указывают на время происходящего. Например, "нынешняя ситуация" (Date:abs + _N_)

2.7 Одиночные именованные сущности и анафора

Анафорические упоминания **Organization**, **GPE**, **Location** и **Facility** выделяются как именованные сущности, если обозначают конкретную сущность, полное имя которой зачастую упоминается в контексте.

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 **Organization**. Анафор “оппозиция” также выделяется как именованная сущность, так как ссылается на тот же конкретный объект.

При разметке кореференции выражение “оппозиция” следует выделить как анафор.

Примеры выделения одиночных анафорических сущностей **Organization**, **GPE**, **Location**, **Facility** и **Event** показаны на Рисунке 53.



Рис. 53: Одиночные **Organization**, **GPE**, **Location** и **Facility**

Подробнее о выделении одиночных **Event** – в Разделе 2.5.5.

Примечание. Одиночные слова, не указывающие на конкретный объект, не выделяются как сущность.

3 Разметка кореференции в brat

Упоминания (меншны) в **brat** отмечаются метками X, Y, X-ATTR и DEF, в зависимости от того, каким типом связи они будут связаны с другими упоминаниями. На Рисунке 54 показаны выделенные упоминания в тексте.

Все выделенные упоминания нужно попарно связать кореферентными связями. Всего в **ru:corner** четыре вида кореферентных связей: IDENT, ANAPH,

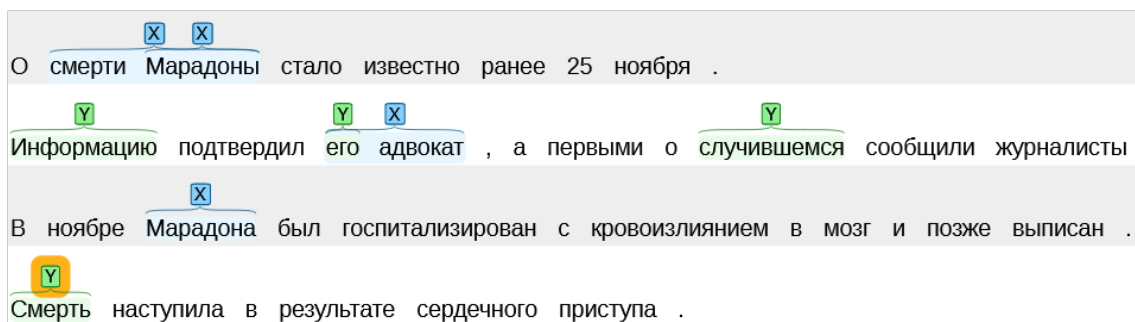


Рис. 54: Пример выделения упоминаний в brat

`CATAPH` и `APPOS`. Кроме того, предусмотрена связь `ELLIPSIS`, с помощью которой неполные упоминания дополняются и могут участвовать в `IDENT` цепочках. Подробную информацию об этих связях можно найти в Руководстве по разметке кореференции в `ru:corner`.

На Рисунке 55 показан пример текста с размеченными связями между упоминаниями.

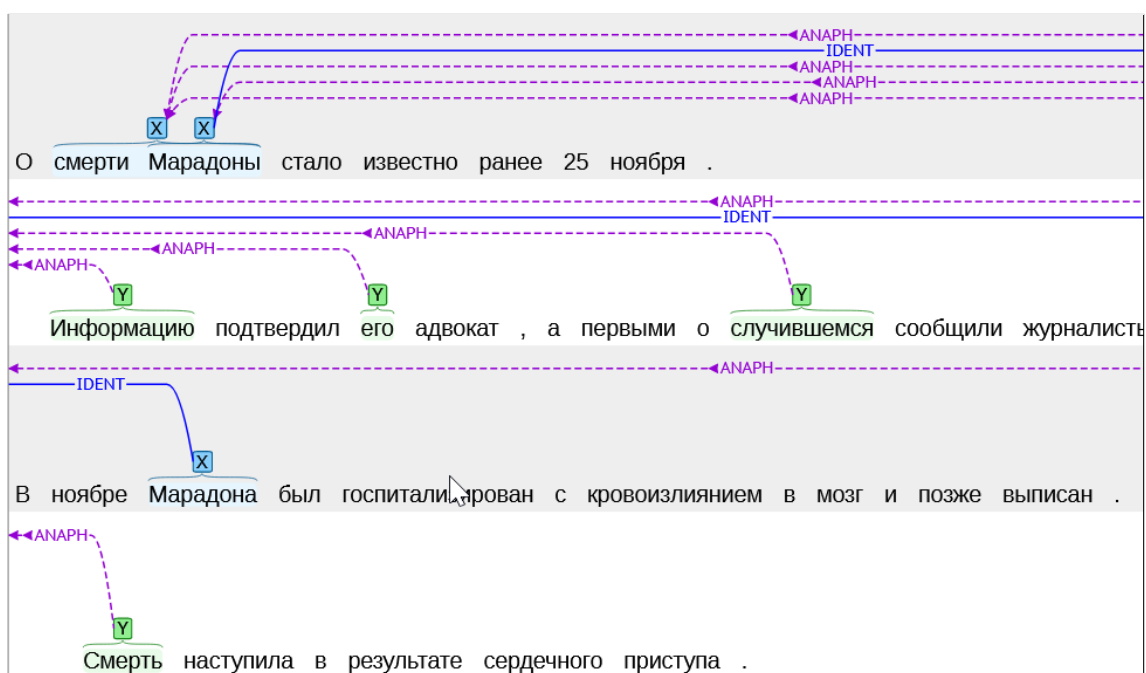


Рис. 55: Пример кореферентной разметки в brat

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

Связь `IDENT` ненаправленная, связь `ANAPH` всегда направлена влево, связь `CATAPH` всегда направлена вправо, а связи `APPOS` и `DEF` могут быть направлены и влево, и вправо.

Для выбора другой метки связи, выберите её из списка (56).

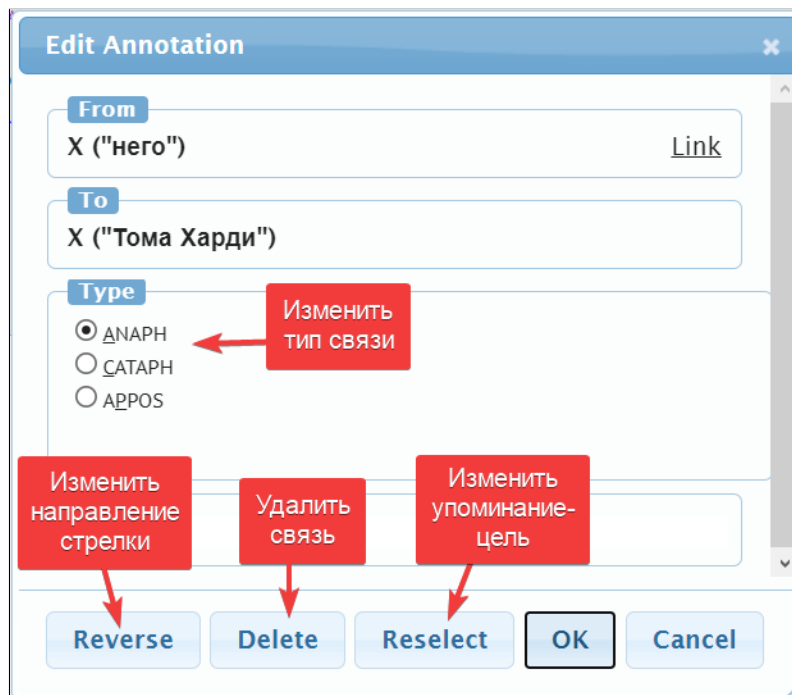


Рис. 56: Изменение или удаление связи между упоминаниями

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.

Примечание. Важно изначально правильно разметить типы упоминаний \boxed{X} , \boxed{Y} , $\boxed{X-ATTR}$ и \boxed{DEF} , так как разные пары упоминаний можно связать только подходящим типом связи.

- IDENT можно протянуть только от \boxed{X} к \boxed{X}
- ANAPH/CATAPH - $\boxed{Y} \rightarrow \boxed{X}$
- APPOS - $\boxed{X-ATTR} \rightarrow \boxed{X}$
- ELLIPSIS - $\boxed{DEF} \rightarrow \boxed{X}$

4 Разное

4.1 Особенности разных жанров

В `ru:corner` включены тексты из разных жанров, каждый из которых содержит коллекции текстов из одного или нескольких открытых интернет-источников. Для всех жанров действуют одинаковые правила разметки, описанные в Таблице 1 и Разделе 2.

Ниже перечислен полный список жанров и некоторые их особенности.

1. Новости (newswire)

Тексты новостей, как правило, написаны в новостном, публицистическом стиле, с небольшим количеством грамматических ошибок. Текст содержит новость целиком или же ее начало, если полный текст новости был длинным.

2. Интервью (interview)

В текстах интервью каждый новый абзац – очередная реплика собеседника. Каждый текст – случайный отрывок из диалога.

3. Художественная литература (fiction)

Литературные тексты могут включать в себя разные жанры, в том числе включать в себя тексты только для взрослой аудитории (18+) и нецензурную лексику. Все тексты литературного жанра следует размечать наравне с остальными. Целиком как `!!!INVALID DOCUMENT!!!` размечаются тексты, преимущественно содержащие в себе стихотворения или написанные на иностранном языке.

В литературных текстах часто встречаются имена персонажей. Чаще всего эти персонажи выдуманные. Тем не менее, при разметке этих имен, атрибут `Ficiton` следует ставить только тогда, когда имя персонажа нетривиально, и широко известно, что это вымышленный персонаж. Так, например, имена вроде *“Виктор, Татьяна, Иван Петрович, Владимир Путин”* и подобные тривиальные имена, которые часто встречаются в реальной жизни, следует выделить как `Person` без атрибута `Fiction`. В свою очередь, упоминания вроде *“Штирлиц, Бэтмен, Вовочка”* и т.п., которые известны как вымышленные персонажи, следует отметить атрибутом `Fiction`.

4. Объявления (advert)

В категории объявлений находятся продающие тексты. Зачастую они содержат в себе описание продаваемого товара или услуги, а также список его характеристик.

Эмодзи (смайлики) заменены на тег `[emo]` для корректного отображения в `brat` – размечать этот тег не нужно.

Размеры товаров в числовом или буквенном обозначении, например, “50, 44-46, XL” следует размечать как `Ordinal`. Размеры товаров формата “800x1200” следует обозначать как `Quantity`.

Графики работ, например *Пн-Пт, Сб, Вс*, следует выделять как `Time: period`.

Тексты объявлений часто написаны автором с ошибками. В частности, знаки препинания могут быть не отделены от слов. Такие случаи следует всегда отмечать как `!!!INVALID DOCUMENT!!!`.

5. Отзывы (feedback)

6. Соцсети (socials)

7. Пьесы (theatrical)

8. Википедия (wiki)

4.2 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 57). Необходимо переразметить сущности таким образом, чтобы они не пересекались.

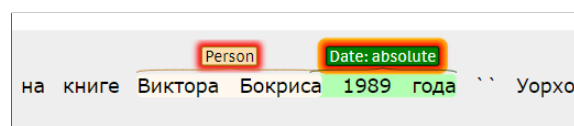


Рис. 57: Пример ошибки при пересечении сущностей.

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

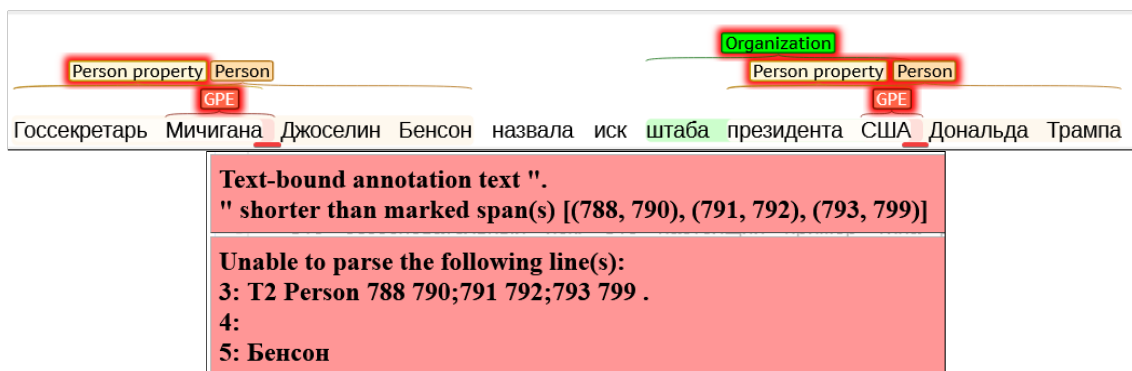


Рис. 58: Лишние пробелы на пересечении сущностей вызывают ошибки.

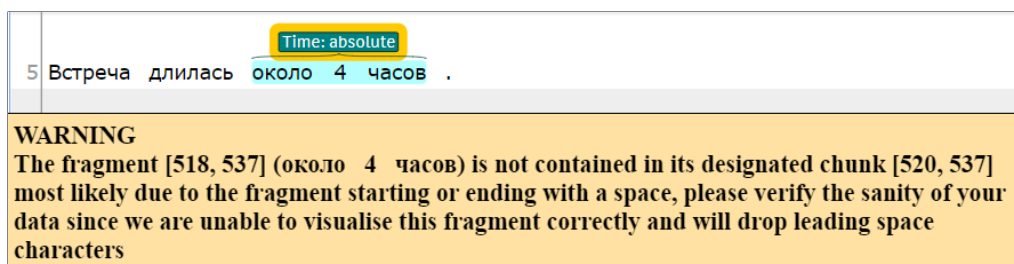


Рис. 59: Пример предупреждения при наличии пробелов в начале сущности.

- **Пробелы в начале сущности.** Если в начало выделенной сущности попали пробелы, высветится предупреждающее сообщение (Рисунок 59). 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.

- **Ошибка токенизации.** Могут встречаться тексты, в которых цельные предложения разбиты на несколько строк, или, наоборот, на одной строке находится два или более предложений. Такие случаи размечаются как **!!!INVALID DOCUMENT!!!**, как описано в разделе 1.5.

4.3 Как быстро выделить текст без лишних пробелов

NB! Инструкция ниже актуальна только для браузеров на движке Chromium (Chrome, Opera и др.).

При работе в Mozilla Firefox, инструкция не работает.

Поскольку между всеми словами в **brat** три пробела, при выделении двойным кликом одиночных сущностей, в него также войдут пробелы в конце сущности.

Для быстрого выделения сущностей без лишних пробелов в конце выполните следующие шаги:

1. Дважды кликните и удерживайте курсор на одном слове.
- 2' Для выделения сущности из одного слова, слегка сместите курсор влево.
- 2'' Для включения нескольких слов в сущность без лишних концевых пробелов, всё так же удерживая мышку, сместите курсор вправо до последнего слова.
3. Выберите тип сущности. В результате в конце сущности не будет лишних пробелов.

Вышеописанные шаги изображены на Рисунке 60.



Рис. 60: Быстрое выделение сущностей без лишних пробелов в конце

A Changelog

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 сущностей.
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.

03.03.2021

1. Исправлены некоторые формулировки и некоторые скриншоты после ревью.

04.03.2021

1. Удален неактуальный раздел 2.1.2 (про окно в 3 предложения). Нужная информация перенесена в соседние разделы.
2. Постредактура: исправлены некоторые формулировки, размеры скриншотов, интервалы и т. п.
3. **Обновлена Таблица 1.** Исправлены некоторые примеры, добавлены ссылки на разделы про время
4. Добавлены уточнения с примерами к атрибутам LCont и RCont, а также примеры в Past-UNK и UNK-Fut в раздел 2.6.2.
5. В разделе 2.6.4 добавлена сноска про то, что слова типа “текущий, нынешний” и т. п. выделяются как **Ordinal** только внутри дат. Вне дат они могут быть **Date:abs** с подходящим атрибутом времени.

05-06.04.2021

1. **Обновлена Таблица 1.** В Time:period добавлены примеры “ежедневный, ежемесячный, ежегодный”. (05.04.2021)
2. **Обновлена Таблица 1.** В Time:period добавлен пример “каждый день”. В GPE:Adj добавлен пример “столичный”. (05.04.2021)

08.04.2021

1. **Обновлена Таблица 1.** Добавлены новые примеры.

26.04.2021

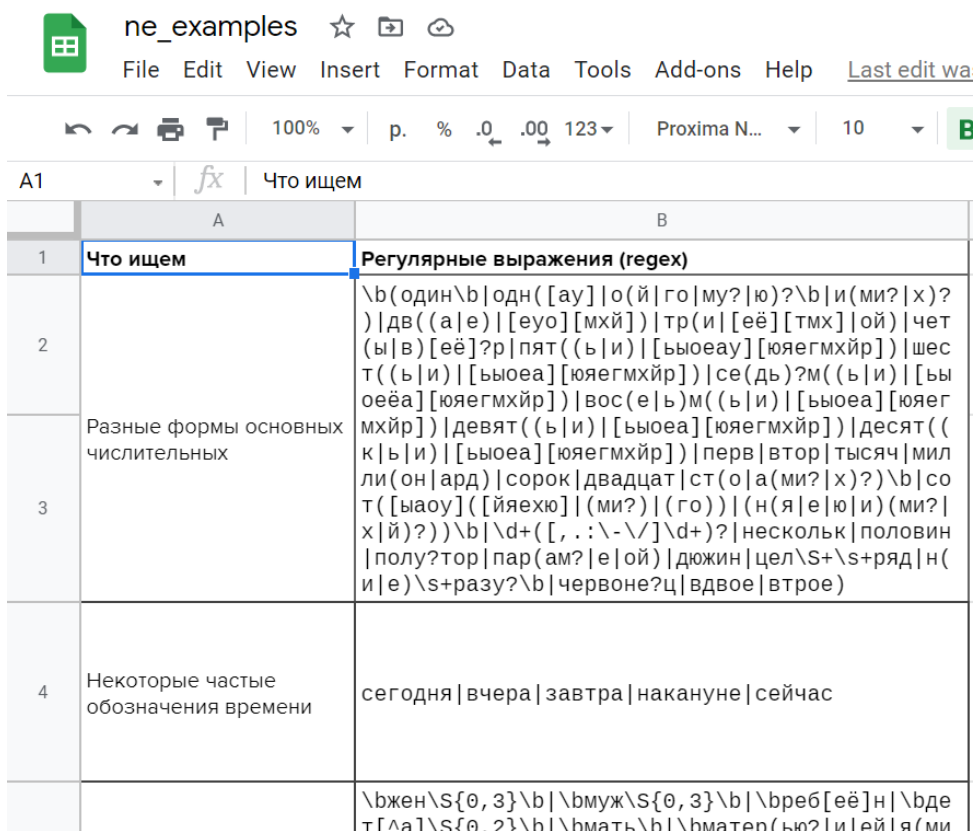
1. **Обновлена Таблица 1.** Добавлены новые примеры, отредактировано описание некоторых категорий, в том числе INVALID DOCUMENT.

28-29.04.2021

1. **Обновлена Таблица 1.** Добавлены новые примеры, отредактировано описание некоторых категорий. (28.04.2021)
 2. **Обновлена Таблица 1.** В описание и примеры **Quantity** добавлено упоминание о фразах с *“целый ряд”*. (29.04.2021)
- 05.05.2021
1. **Добавлено Приложение В** по проверке аннотации в brat.
- 18.05.2021
1. **Обновлена Таблица 1.** Уточнены некоторые описания категорий. Удален атрибут **Plural** из категории **Law**.
- 19.05.2021
1. **Выделен Раздел 4. Добавлен Раздел 4.1.** Описаны жанры текстов, содержащихся в **ru:corner** и их особенности.
 2. **В Раздел 2.6 добавлено уточнение** об указателях *“этот, данный”* и т.п. Уточнение выделено в рамку.
- 24.05.2021
1. **В Разделе 4.1 добавлен абзац про разметку имен в жанре Fiction.**
- 24.05.2021
1. **Обновлена Таблица 1.** В **GPE:adj** добавлен пример *“отечественный”*.
- 13.07.2021
1. **Обновлена Таблица 1.** В примеры **Quantity** добавлены *“дуэт, трио, квартет”*. Добавлены примеры в **Org**.
 2. **Добавлен Раздел 4.3 с иллюстрациями** про быстрое выделение сущностей без пробелов.
 3. **Добавлен Раздел 2.5.2 с иллюстрациями** про выделение **Person** без имени, с вложенными сущностями и без.
 4. **Добавлен Раздел 2.5.5 с иллюстрациями** про особенности выделения событий **Event**.
 5. **Переписан Раздел 2.7, добавлены иллюстрации.** Теперь одиночные упоминания-анафоры некоторых типов сущностей могут выделяться как именованные сущности.
 6. **Исправлено расположение картинок по всему тексту.**
 7. **Исправлены некоторые формулировки по всему тексту.**
- 14.07.2021
1. **Обновлена Таблица 1.** От **Person:Unconscious** отделена категория уникальных объектов (корабли и др.) в новый атрибут **Product:Unique**.
 2. **Обновлен Раздел 2.3.** Добавлен атрибут **Unique** и скриншот выбора атрибутов (Рисунок 21).
 3. **Исправлено расположение картинок и некоторые формулировки.**
- 21.07.2021
1. **Дополнен Раздел 2.5.5** про разметку **Event**, добавлены скриншоты, формализовано выделение событий без названий.
 2. **Актуализирован Раздел 3** про разметку кореференции, заменены или удалены нерелевантные скриншоты.

В Проверка аннотации поиском в brat

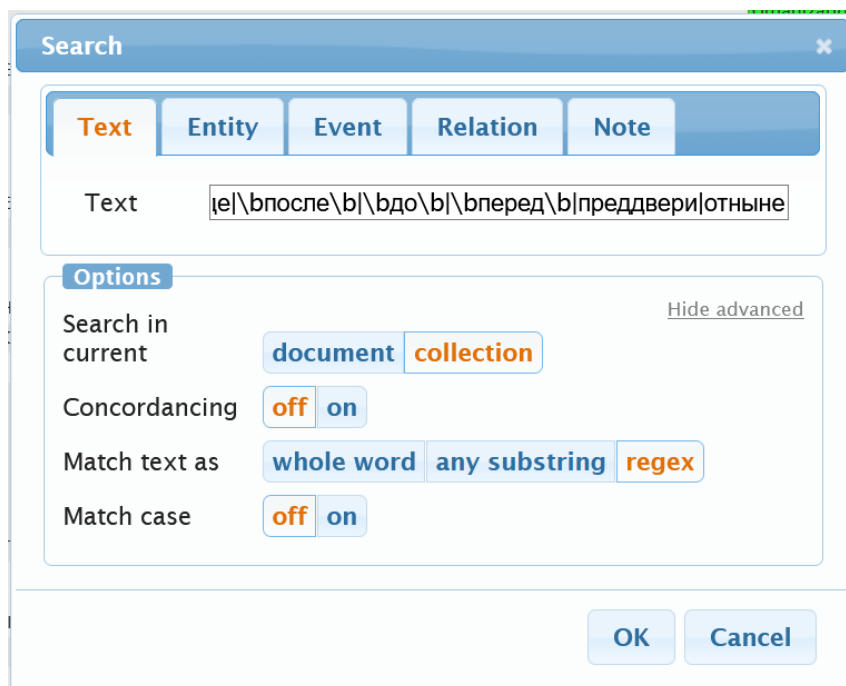
В этой инструкции наглядно описано, как применять регулярные выражения из Google-таблицы (Рисунок 1) для поиска ошибок/пропусков аннотации.



ne_examples		
File Edit View Insert Format Data Tools Add-ons Help Last edit wa		
100% p. % .0 .00 123 Proxima N... 10 B		
A1	fx	Что ищем
	A	B
1	Что ищем	Регулярные выражения (regex)
2	Разные формы основных числительных	<code>\b(один\b один([ау] о(й го му)? ю)?\b и(ми? х)? дв((а е) [еуо][мхй]) тр(и [её][тмх] ой) чет(ы в [её]?р пят((ь и) [ьуюеау][юяегмхйр]) шес(т((ь и) [ьуюеау][юяегмхйр]) се(дь)?м((ь и) [ьуюеау][юяегмхйр]) вос(е ь)м((ь и) [ьуюеау][юяегмхйр]) дев(я(т((ь и) [ьуюеау][юяегмхйр]) деся(т((к ь и) [ьуюеау][юяегмхйр]) перв втор тысяч милли(он ард) сорок двадцат ст(о а(ми? х)?)\b со(т([ьаоу]([йяею] (ми? (го) н(я е ю и)(ми? х й)?))\b \d+([, . : \- \\/]\d+)?) нескольк половин полу?тор пар(ам? е ой) дюжин цел\S+\s+ряд н(и е)\s+разу?\b червоне?ц вдвое втрое)</code>
3		
4		
	Некоторые частые обозначения времени	сегодня вчера завтра накануне сейчас
		<code>\bжен\S{0,3}\b \bмуж\S{0,3}\b \bреб[её]н \bде(т г а л)\S{0,2}\b \bмать\b \bматер(ью? и ей я ми)</code>

Рисунок 1. Таблица в Google с примерами регулярных выражений

Скопируйте одно из регулярных выражений из таблицы и вставьте его в поиск (Рисунок 2).



Search

Text Entity Event Relation Note

Text `|е|\бпсле\b|\бдо\b|\бперед\b|преддвери|отныне`

Options

Search in current

Concordancing

Match text as

Match case

OK Cancel

Рисунок 2. Настройка поиска

Допустим, найдем указатели на время/даты/периоды с помощью регулярного выражения
период|врем|момент|\bv\s+течени|протяжении|\bv\s+ходе\b|\bпосле\b|\bдо\b|\bперед\b|преддвери|отныне

Поиск находит довольно большое количество подходящих случаев (Рисунок 3).

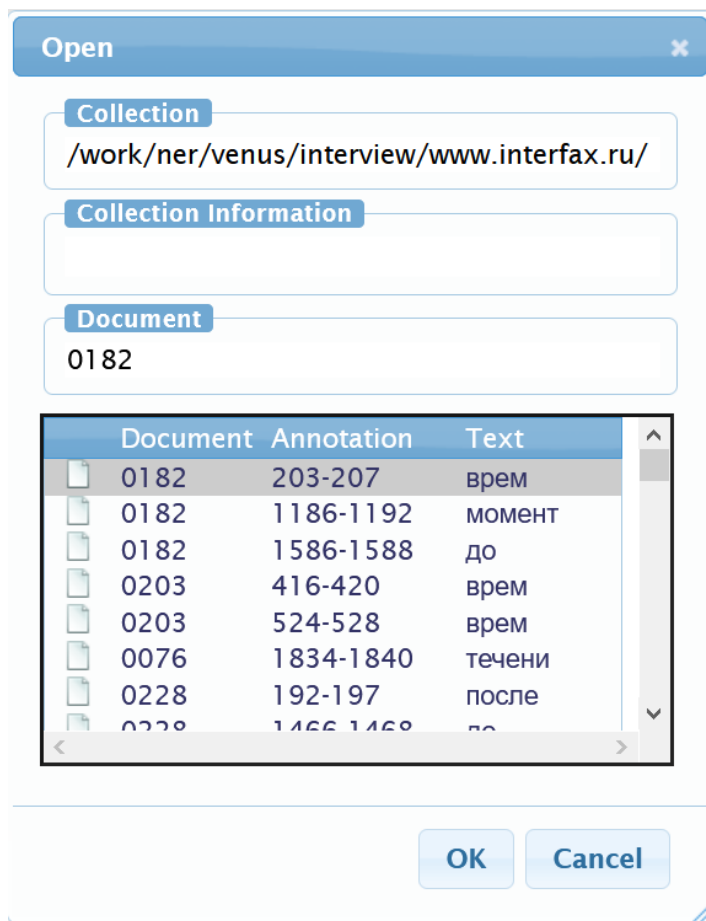


Рисунок 3. Результат поиска

В одном из найденных случаев (Рисунок 4) фразу «*после Конференции*» нужно выделить как **Date: absolute + Past**, поскольку дату события можно восстановить, зная дату конференции.

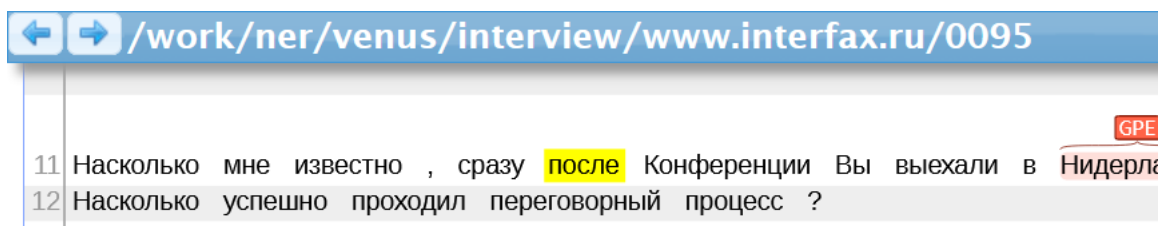


Рисунок 4. Результат поиска

После того, как пройдемся по всем найденным случаям, ищем следующее регулярное выражение. Например, сегодня|вчера|завтра|накануне|сейчас.

Поиском получается найти сущности, которые были пропущены при первоначальной разметке.

1 Для группы "Московская биржа", в которую входит НРД, в этом году главный проект - это введение единого пула обеспечения, биржа сейчас

(Date: absolute + Present)

11 Накануне турецкого заседания "спикер в изгнании" Джемаль Хашмет заявил

(Date: relative + Past)

После прохождения по всем найденным документам и исправления разметки, где это необходимо, переходим к следующему регулярному выражению.

Введем в поиск регулярное выражение для поиска терминов семьи:

\бжен\{0,3}\б|\бмуж\{0,3}\б|\бреб[её]н|\бдет[^\a]\{0,2}\б|\бмать\б|\бматер(ью?|и|ей|я(ми?|х))\б|\боте?ц(а(ми?|х)|е
 |ы|у|о(м|в)?)\б|\ботче\б|\ббрат|\бсестр|\бпап\{0,3}\б|\бмам\{0,3}\б|\бснох\{0,3}\б|\бдевер|\бсвояк|\бкум\{0,3}\б|\бсемь\{1,
 3\}\брод\{0,3}\б|\бродственни|\бдяд\{1,3\}\бтет\{1,3\}\бвневест|\бжени(х|ш)|\бсын|\бдоч(ь|а|ер(и|ью?|я(ь|ми)?|ях|ей)|\бпе
 рвен|\ббаб|\бдед|\бправн|\бпра[бд]|\бвну[кч]|\бплемянн|\бзоловк|\бтеш\{1,3\}\бтест[^\aоы]\{0,3}\б|\бсвекр|\бзят\{1,3\}\b
 \бпадчериц|\бпасын|\бшурин|\бсвоя[кч]|\бдевер|\бсупруг|\бсирот

В тексте ниже получилось найти и доразметить термины «сирота», несколько «ребенков» и «родители». Все вхождения (размеченные и пропущенные) подсвечиваются желтым.

36 И когда выявляется **ребенок-сирота** , органы образования , здравоохранения , соцзащиты , **МВД** делают все , чтобы не помещать его в детский дом , а устроить в уже подготовленную **семью** .

38 Кстати , сколько в **России** **сирот** ?

39 В **последние** **годы** звучит цифра **700** **тысяч** ...

41 В нашей статистике даже те **дети** , которые уже устроены в **семьи** , считаются **сиротами** .

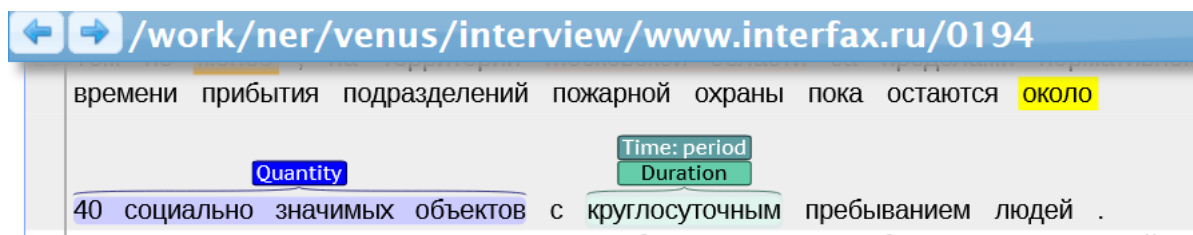
42 У меня лично язык не поворачивается назвать **ребенка** **сиротой** , если он устроен в **семью** .

43 Это даже обидно и для **ребенка** , и для **семьи** , которая его взяла .

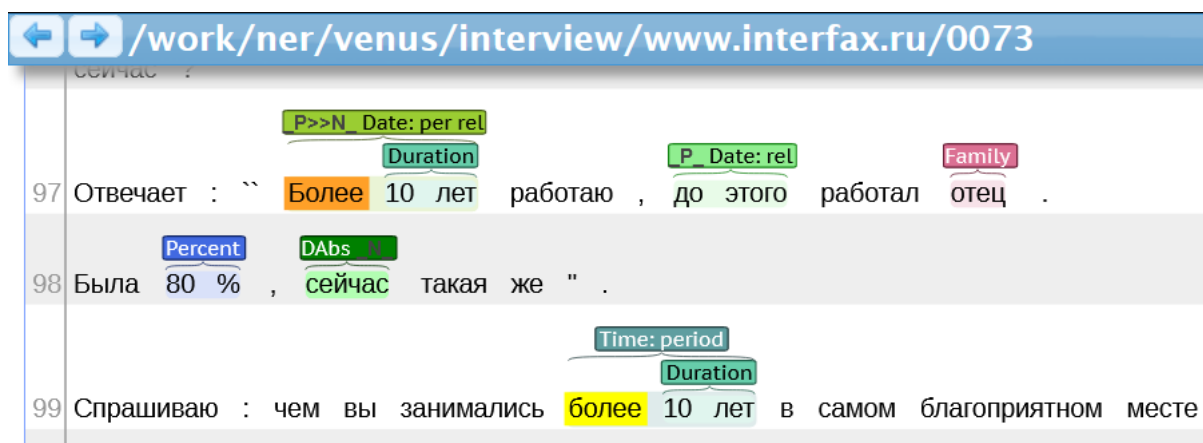
44 В федеральном банке данных число **детей-сирот** , которые не устроены в **семьи** , стабильно снижается

Поиск регулярного выражения \b(более|менее|ещё|еще|свыше|почти|около|больше|меньше)\b

позволяет найти и доразметить пропущенные конкретизаторы.



В примере ниже поиском нашлись два рядом стоящих случая, когда слово «более» должно быть включено в сущность **Duration** («10 лет» не равно по длительности «более 10 лет»).



Чтобы поиск выдавал меньше шума и больше полезных результатов, иногда регулярные выражения в Google-таблице могут обновляться. Поэтому рекомендуется при проверке копировать эти «регулярки» напрямую из этой таблицы.