

Оригинальная статья / Original Article

DOI: 10.31857/S160578800023677-0

Лексикализованные обозначения времени: опыт корпусной разметки

© 2022 г. С. П. Тимошенко

Научный сотрудник лаборатории компьютерной лингвистики
Института проблем передачи информации им. А.А. Харкевича (ИППИ РАН),
Россия, 127051, Москва, Большой Каретный переулок, д. 19, стр. 1
timoshenko@iitp.ru

Резюме. В изучении обозначений времени имеется перекося в сторону изучения грамматических категорий времени и вида, поскольку в большинстве языков эти категории присутствуют и принимают на себя основной груз “работы” по передаче временных значений. Однако ту же информацию можно передавать и лексическими средствами. Более того, каталогизация именно лексических средств обозначения времени и сравнение систем этих средств в разных языках позволяют очертить границы человеческого восприятия времени, понять, что в нем универсально, а что — специфично. Мы ставим перед собой цель провести корпусное исследование именно лексических средств обозначения времени в русском языке. Для этого мы разработали формат временной разметки и выполнили пилотную аннотацию. В статье описывается использованный формат и даются лингвистические обоснования решений, принятых при его разработке.

Благодарность. Автор благодарит Министерство науки и высшего образования РФ за финансовую поддержку работы (грант № 075-15-2020-793 “Компьютерно-лингвистическая платформа нового поколения для цифровой документации русского языка: инфраструктура, ресурсы, научные исследования”).

Ключевые слова: семантическая разметка, временная разметка, темпоральная разметка, SynTagРус, TimeML, лексикализованные обозначения времени, временная локализация.

Для цитирования: Тимошенко С.П. Лексикализованные обозначения времени: опыт корпусной разметки // Известия Российской академии наук. Серия литературы и языка. 2022. Т. 81. № 6. С. 70–88. DOI: 10.31857/S160578800023677-0

Lexical Expressions with Temporal Semantics: the Corpus Study

© 2022 Svetlana P. Timoshenko

Research Associate at the Laboratory of Computational Linguistics
of the Kharkevich Institute for Information Transmission Problems (IITP RAS),
19 Bld.1 Bolshoy Karetny Lane, Moscow, 127051, Russia
timoshenko@iitp.ru

Abstract. The field of study that can be named “temporal semantics” is marked by the domination of grammatical studies. The researchers’ interest is focused on the tense and the aspect, lexical expressions with temporal meaning stay in the background. Obviously, it is grammatical means that for the most part express temporal meanings: they are present in most languages and in every sentence in these languages. However, lexical means can convey the same information. We argue that the cataloging of lexical means and comparing the systems of these means in different languages allow us to outline the boundaries of human perception of time and to better understand what is universal and what is specific in it. Our goal is to conduct a corpus study of temporal lexical expressions in the Russian language. To do this, we developed an annotation scheme and performed a pilot markup on some fragments of the SynTagРус. The paper describes our annotation scheme and provides linguistic grounds for the decisions made during its development. We

also compare our annotation scheme with the TimeML scheme, the most used annotation scheme for temporal information in the world. We show that despite the elaborated structure and a large set of elements, TimeML fails to grasp some important fragments of the expressions' semantics. The proposed annotation scheme is more expressive and flexible and suits for the annotation of the complex examples: *sekretarša na minutu vyšla* ('the secretary left for a minute'), *delat' čto-libo tretij čas podryad* ('do something more than two hours in succession' — literally 'for the third hour in a row').

Acknowledgements. The author thanks the Ministry of Science and Higher Education of the Russian Federation for financial support of the work (grant No. 075-15-2020-793 "A new generation computer-linguistic platform for digital documentation of the Russian language: infrastructure, resources, scientific research").

Key words: semantic markup, temporal markup, SynTagRus, TimeML, temporal expressions, temporal localization, annotation scheme.

For citation: Timoshenko, S.P. *Leksikalizovannye oboznacheniya vremeni: opyt korpusnoj razmetki* [Lexical Expressions with Temporal Semantics: the Corpus Study]. *Izvestiâ Rossijskoj akademii nauk. Seriâ literatury i âzyka* [Bulletin of the Russian Academy of Sciences: Studies in Literature and Language]. 2022, Vol. 81, No. 6, pp. 70–88. (In Russ.) DOI: 10.31857/S160578800023677-0

Введение

Мы хотим исследовать, как человек с помощью естественного языка передает информацию о времени. Мы считаем, что для подобного исследования необходим корпус, содержащий в эксплицитном виде информацию о функционировании темпоральных выражений. Созданию корпуса предшествует разработка формата разметки.

Лингвистические исследования знают много успешных примеров применения корпусного подхода, когда новое знание достигается путем разметки в корпусе какого-либо явления. Формат разметки в этом случае представляет собой инструмент исследования. Он может быть основан на каких-либо теориях, описывающих рассматриваемое явление или, наоборот, иметь эмпирический характер. В качестве успешных примеров можно привести PropBank и Penn Discourse TreeBank (далее — PDTB). PropBank — это ресурс, созданный для исследования варьирования в синтаксическом оформлении семантических актантов при глаголах (диатезы в широком смысле) [1]. PDTB — ресурс для изучения дискурсивных структур [2]. Первым шагом в их разработке было выполнение соответствующей разметки на одной и той же коллекции текстов — на корпусе Wall Street Journal (WSJ) объемом 1 миллион слов. Этот корпус также имеет синтаксическую разметку в виде деревьев зависимостей, что дает исследователям возможность анализировать взаимосвязи между синтаксическими, семантическими и дискурсивными структурами, а также разрабатывать и оценивать практические приложения, использующие разные комбинации синтаксических, семантических и дискурсивных признаков.

Существенно, что и в случае разметки семантических аргументов глаголов, и в случае разметки дискурсивных связей формат разметки задавался не столько "сверху", на основе имеющихся теоретических наработок, сколько "снизу", на материале данных. Авторы PDTB прямо называют свой подход "теоретически нейтральным". Что же касается PropBank'a, то его авторы отказались от какого-либо заданного набора семантических ролей и приписывали их каждому глаголу по отдельности. Таким образом, помимо собственно размеченного корпуса, результатом проекта явился большой словарь английских глаголов, содержащий информацию об их валентных структурах.

Разумеется, уже предпринимались попытки исследовать подобным образом временные выражения. Мы коротко опишем их в первом разделе статьи. Самый влиятельный результат в этой области достигнут коллективом под руководством Джеймса Пустейовски. Этот коллектив разработал формат разметки TimeML. Его описание читатель также найдет в первом разделе. Формат этот не свободен от недостатков. В частности, лингвистическая работа с материалом в этом формате сопряжена с заметными неудобствами. Поэтому мы приняли решение разработать новый формат. Во втором разделе статьи очерчивается круг явлений, на который мы ориентировались при разработке формата. В третьем разделе мы кратко описываем предлагаемый формат и показываем, как применять его к образцовым случаям. Четвертый раздел посвящен трудным случаям, с которыми мы столкнулись в процессе сплошной разметки текста. Для их описания мы также предлагаем решения.

1. Краткий очерк истории темпоральной разметки

Большинство проектов, ставивших своей целью создание формата разметки для временной информации, связано отношением преемственности. Основные результаты в этой области достигнуты в рамках трех исследовательских инициатив: “Message Understanding Conference” (MUC, “Конференция по вопросам понимания сообщений”) [3], ее наследницей “Translingual Information Detection, Extraction and Summarization” (TIDES, “Многоязычное обнаружение, извлечение и обобщение данных”) [4] и “The Automatic Content Extraction” (ACE, “Автоматическое извлечение содержания”) [5]. Все они были связаны с практическими задачами разработки приложений и оценки их эффективности.

Впервые задача определения абсолютных временных выражений (absolute time expressions) была поставлена перед участниками MUC-6 в 1995 году как подзадача извлечения именованных сущностей. Через два года, в рамках соревнования MUC-7, в рассмотрение были добавлены относительные выражения (relative time expressions). Соотнесение выражений с событиями и установление временных связей в задачу не входило. Именно тогда появилась разметка в формате XML, где соответствующий выражению фрагмент текста заключался в тег TIMEX и получал одну из двух возможных классифицирующих помет — DATE или TIME.

<TIMEX TYPE="TIME">twelve o'clock noon</TIMEX> ‘двенадцать часов пополудни’

<TIMEX TYPE="DATE">third quarter of 1991</TIMEX> ‘третий квартал 1991 года’

Спецификация TIDES [6] наследует принципы, предложенные MUC, и рассматривает временные выражения сами по себе, вне связи с событиями, однако делает следующий шаг по пути к эксплицитному представлению их семантики. Основной тег разметки, TIMEX2, обладает большим количеством атрибутов. Специфическим образом очерчивается круг выражений, подлежащих разметке: выбираются группы слов и выражений, которые легко соотнести с осью времени. Слова и выражения, чья семантика признается “размытой” (vague), из рассмотрения исключаются. Формат следующего поколения, TimeML, полностью унаследовал этот подход к временным выражениям. TimeML также является самым распространенным форматом разметки и заслуживает подробного описания, которое будет дано ниже. Сразу отметим, что главное новшество TimeML по сравнению с предшественниками — это

одновременная разметка временных выражений и обозначений событий, а также установление временных отношений. Временные отношения могут связывать два времени, два события или время и событие.

TimeML, однако, не свободен от недостатков, поэтому исследователи продолжают разрабатывать новые форматы разметки. Например, в 2021 году был предложен формат TIE-ML (Temporal Information and Event Markup Language) [7]. В качестве основного недостатка TimeML, побудившего к разработке альтернативы, авторы указывают его сложность. Поэтому их предложение отличается радикальной простотой. Размечаются только события. Текстовым коррелятом события считается клауза. Сведения о времени, которые можно извлечь из грамматической формы, кодируются в соответствии со схемой, предложенной Райхенбахом [8] и основанной на трех типах временных переменных: времени события (Event Time), времени отсчета (Reference Time) и времени речи (Speech Time). Временные выражения не размечаются. Информация, которую они передают, отражается разметкой в качестве атрибута события и только в том случае, если представляет собой конкретный временной “якорь”, например

<s> <c reference="264 BC"> The First Punic War broke out on the island of Sicily in 264 BC. </c> </s>

Размечено событие — начало Пунической войны. Ему соответствует клауза, совпадающая с предложением. Временное выражение отражается с помощью атрибута reference.

TimeML

Поскольку разработка TimeML была мотивирована более широкими, лингвистическими задачами компьютерной науки, формат охватывает только события и временные промежутки, чье положение на временной оси можно однозначно определить. Временные характеристики в генерических высказываниях, например, *Евреям запрещено убивать друг друга*, и в описаниях типичных паттернов деятельности, например, *Господин Сунуну регулярно пользуется корпоративными самолетами для политических поездок* (примеры взяты из [9, с. 7]), выходят за пределы области применения TimeML. Кроме того, от формата TIDES наследуется принцип, согласно которому не размечаются обозначения времени с “размытой” семантикой, которые трудно соотнести с осью времени. В стандарте TIDES [6] вводится понятие “не-триггера” (non-trigger), то есть не подлежащего

разметке выражения: “не-триггеры”, несмотря на свою временную семантику, — это класс выражений, чье расположение на временной оси трудно определить. Эти “не-триггеры” подразделяются на несколько типов (мы не снабжаем примеры из инструкции пословным переводом, поскольку перевод их часто определяется контекстом. Вместо этого мы приводим перечень аналогичных единиц русского языка, которые при определенных условиях выступают переводными эквивалентами одной или нескольких исходных английских единиц):

- упорядочивающие выражения и выражения, задающие последовательность: *ahead, next, subsequent, later, preceding, previous, previously, early, earlier, heretofore, so far, afterwards, before, after, beforehand, following, eventually, finally, [for] the first time, first, later, originally, previously, subsequently, then, already, yet, still, ever*. Русскоязычные аналоги: *после, позже, предшествующий, предыдущий, грядущий, будущий, наступающий, раньше, ранее, прежде, до сих пор, с тех пор, до, в конце концов, наконец, в первый раз, сначала, первоначально, впоследствии, потом, уже, еще, заранее, сначала, некогда*;
- адвербиалы, отвечающие на вопросы “как долго / как быстро?”, “как скоро?": *immediately, momentarily*. Русскоязычные аналоги: *немедленно, мгновенно, через минуту и т.д.*;
- обозначения длительности, не поддающиеся количественной оценке: *longstanding, long-term, long, permanently, temporarily*. Русскоязычные аналоги: *долго, долгий, долгосрочный, постоянный, временный*;
- выражения с отрицательной семантикой и обозначения несуществующих времен. Этот тип иллюстрируется двумя примерами: *Five-time champion Rick Swenson wasted no time moving to the front of the pack in the Iditarod Trail Sled Dog Race*. ‘В Айдитародской гонке на собачьих упряжках пятикратный чемпион Рик Свенсон, не мешкая (дословно ‘потратил никакое время’), вырвался вперед’. Русскоязычным аналогом можно считать выражение: *ни минуты (не теряя)*. *No date has been set for another summit meeting*. ‘Дату следующей встречи на высшем уровне не назначили’. Дословно ‘никакая дата не назначена’;
- обозначения частотности и кратности без указания на период времени: *once, frequent, frequently, too often, usually, always, three times, on five occasions, normally, commonly, generally, constantly, often, repeatedly, sometimes, rarely,*

never. Русскоязычные аналоги: *часто, частый, обычно, обыкновенно, все время, снова и снова, редко, иногда, никогда, дважды, X раз*. Однако в случае, если подобная единица входит в состав выражений, обозначающих периодичность, то она подлежит разметке. Например, выражение *twice a week* ‘дважды в неделю’ считается единой сущностью класса TIMEX3, т.е. слово *дважды* в нее включено;

- обозначения времени в составе имен и названий: *роман Оруэлла “1984”*.

Теги, отмечающие фрагменты текста, таковы: EVENT (‘события’), TIMEX3 (‘временные выражения’), SIGNAL (‘сигналы’). Все упоминания ситуаций, которые случаются или происходят, получают тег EVENT. Все релевантные обозначения времени — относятся к классу TIMEX3 (цифра 3 означает преемственность формата относительно TIDES и MUC). В некоторых случаях тег TIMEX3 не соотносится ни с каким фрагментом текста, обозначая подразумеваемое время. В тег SIGNAL заключаются вспомогательные слова временного выражения, главным образом предлоги.

Для хранения информации о связях используются теги, которые записываются после текста и “опираются” на индивидуальные сущности типа EVENT, TIMEX3 и SIGNAL.

Еще один тег занимает промежуточное положение — он не маркирует текстового фрагмента и не обозначает связи между маркированными фрагментами. Это тег MAKEINSTANCE. Элементы xml с таким тегом порождаются отдельно и обозначают экземпляры событий, обозначаемых выражениями, заключенными в тег EVENT. Так, в предложении *John taught on Monday and Tuesday* ‘Джон давал уроки в понедельник и во вторник’ тег EVENT относится к глаголу *taught*, но поскольку временных обозначений два, то порождается также два “экземпляра” события, производных от *taught* и связанных каждый со своим днем недели. Экземпляры событий существуют в виде элементов MAKEINSTANCE.

Формат TimeML предусматривает следующие типы связей между элементами типа MAKEINSTANCE, TIMEX3 и EVENT: аспектуальная связь, подчинительная связь и временная связь. Каждому типу связи соответствует отдельный тег. ALINK (Aspectual Link ‘аспектуальная связь’) используется, когда в тексте присутствует указание на фазу ‘события’, например, *started to read* ‘начал читать’. SLINK (Subordination Link ‘подчинительная связь’) используется для

разметки большого количества более или менее грамматических значений, которые в английском языке выражаются отдельным глаголом: модальные значения (*wanted to buy* ‘хотел купить’), эвиденциальные значения (*John said he bought* ‘Джон сказал, что купил’), фактические значения (*managed to leave* ‘удалось уйти’) и т.д. Наконец, TLINK (Temporal Link ‘временная связь’), отражает временную связь.

Каждый тег имеет свой набор атрибутов, у некоторых он весьма обширен. Мы не будем перечислять их все, просто продемонстрируем возможности формата на образцовом примере разметки из [9, с. 60].

A major earthquake struck Los Angeles three years ago today.
‘Сегодня ровно три года, как в Лос-Анджелесе случилось сильное землетрясение’

```
A major
<EVENT eid="e1" class="OCCURRENCE" >
  earthquake
</EVENT>
<MAKEINSTANCE eiid="ei1" eventID="e1"
tense="NONE" aspect="NONE" pos="NOUN"/>
<EVENT eid="e2" class="OCCURRENCE" >
  struck
</EVENT>
<MAKEINSTANCE eiid="ei2" eventID="e2"
tense="PAST" aspect="PERFECTIVE"/>
  Los Angeles
  <TIMEX3 tid="t1" type="DURATION" value="P3Y"
beginPoint="t2" endPoint="t3">
    three years ago
  </TIMEX3>
  <TIMEX3 tid="t2" type="DATE" value="2002-07-12"
temporalFunction="true"
anchorTimeID="t0">
    today
  </TIMEX3>
  <TIMEX3 tid="t3" type="DATE" value="1999-07-12"
temporalFunction="true"
anchorTimeID="t1"/>

  <TLINK eventInstanceID="ei1" relatedToEvent="ei2"
relType="IBEFORER"/>
  <TLINK eventInstanceID="ei1" relatedToTime="t3"
relType="IS_INCLUDED"/>
```

В этом предложении выделено два события (EVENT) — earthquake ‘землетрясение’ и struck ‘случилось’. Соответственно, создано две инстанции событий, ei1 и ei2. Они связаны между собой временной связью, которую можно прочесть так: землетрясение ei1 непосредственно предшествовало событию ei2 ‘случилось’ (по определению отношения IBEFORE).

Заметим, что между существительным earthquake и глаголом struck имеется определенное семантическое соотношение, которое можно описать следующим образом: существительное

обозначает событие, глагол выступает сказуемым при этом существительном и имеет прямое дополнение, образуя высказывание. Смысл высказывания: ‘событие имело место, и его главный участник обозначен с помощью прямого дополнения глагола’. Иными словами, основная семантическая нагрузка лежит на существительном-подлежащем, именно оно отвечает на вопрос “что случилось?”, а глагол выполняет две вспомогательные функции — несет на себе грамматические показатели высказывания (время и модальность) и присоединяет к событию его семантический актанта. Аналогичное соотношение можно усмотреть в большом количестве устойчивых словосочетаний: *случилась беда* (с кем-то), *дружба связывает* (кого-то), *коррупция разъедает* (общество). Поэтому такое соотношение является регулярным. В теории лексических функций оно классифицируется как функция FUNC1 [10]. Ошибочно в таких случаях усматривать два события — обозначаемое существительным и обозначаемое глаголом. Перед нами событие, обозначенное с помощью двух слов. При выполнении разметки здесь возможны разные решения: считать обозначением события только существительное и игнорировать глагол (по семантическому критерию), считать обозначением события только глагол и игнорировать существительное (по грамматическому критерию), считать обозначением единого события оба слова (<EVENT> earthquake struck </EVENT>), в формате TimeML считать два разных слова манифестациями одной инстанции (MAKEINSTANCE), на худой конец — считать, что два события, производных от существительного и от глагола соответственно, имели место в одно и то же время. Предлагаемое решение, опирающееся на идею непосредственного следования, кажется нам наименее удачным.

Возвращаясь к рассмотрению примера из спецификации, отметим, что предложение также размечено тремя сущностями типа TIMEX3. TIMEX3 с идентификатором t1 — это выражение *three years ago*, оно обозначает интервал. TIMEX3 с идентификатором t2 — это выражение *today*, дата, являющаяся концом интервала t1. Ее абсолютное значение определяется на основе t0, времени создания/публикации текста. TIMEX3 с идентификатором t3 не имеет эксплицитного выражения в тексте. Он появляется в результате логического вывода: у интервала t1 должен быть и другой конец, помимо t2. Про него известно, что это дата, чье абсолютное значение определяется на основе t1. Вторая временная связь — между землетрясением ei1 и датой t3, состоящая в том, что землетрясение ‘включено’ в указанный день.

Подведем итоги — TimeML не просто сложный, а очень сложный в применении формат. Помимо большого количества элементов, атрибутов и их возможных значений, в которых легко запутаться, у него есть и другие недостатки:

1) Содержательная неполнота. Множество размечаемых выражений ограничено “точными” выражениями, то есть такими, семантика которых относительно просто соотносится с осью времени. Исключаются из рассмотрения именно те выражения, которые в значительной степени отвечают за “связность” временного плана большинства текстов.

2) Искусственная изоляция временной разметки от лингвистического анализа текста. Предполагается, что разметка выполняется по “сырому” тексту, поэтому формат включает в себя много грамматической информации: тег MAKEINSTANCE имеет целых 5 атрибутов для хранения грамматической информации — tense, aspect, nf_morph, polarity, modality.

Таким образом, морфология и даже фрагменты синтаксиса (polarity хранит сведения о наличии соответствующего модификатора) оказываются “влиты” в семантическую по преимуществу разметку, размывается граница между грамматикой и семантикой. Кроме того, непропорционально большая доля тегов, атрибутов и их значений оказывается ориентирована на грамматику английского языка. Например, тег ALINK и его атрибуты совершенно неприменимы к русскому языку с его богатой и запутанной системой видо-временных значений.

Первый недостаток представляется нам самым существенным. Его одного, в принципе, достаточно, чтобы мотивировать разработку нового формата, поэтому мы начали работу с определения круга интересующих нас явлений и исходили из того, что практически все “не-триггеры” TimeML должны в него войти.

2. Круг рассматриваемых явлений

Можно выделить 4 интуитивно понятных типа временной информации: локализация во времени, длительность, периодичность, последовательность (во времени). Локализация — это ответ на вопрос “когда?”, длительность — ответ на вопрос “как долго?”, последовательность — ответ на вопросы “после чего?”, “перед чем?”, периодичность — ответ на вопрос “на каком временном отрезке повторяется?”. Периодичность является частным случаем квантификации событий, однако наш предмет исследования ограничен семантикой времени, поэтому мы приняли решение

квантификацию как таковую из рассмотрения исключить. На практике это означает, что выражение *дважды в неделю* подлежит разметке, а выражение *дважды* само по себе — нет.

Приводя иллюстративный материал для каждого типа информации, мы двигаемся от очевидных случаев к нетривиальным и затем к спорным.

локализация

28 сентября, во вторник, завтра, через 5 минут, на третий день (войны / по приезде), в школьные годы, в священный для мусульман праздник, в детстве, в grippe, во время совершения преступления

Примеры, которые исключает формат TimeML: *потом, раньше, когда-то, когда-нибудь*

Специальные конструкции используются в русском языке для локализации через указание на начало и конец события: *с двух часов <понедельника, 1992 года>, по 5 февраля <пятницу>.*

длительность

четыре часа, несколько лет, ночь (напролет), 22-летнее господство Румынии в Бессарабии, третье столетие подряд

Примеры, которые исключает формат TimeML: *долго, долгий, долгосрочный, некоторое время, неопределенный срок*

периодичность

пять раз в неделю, каждый день, по вторникам

Примеры, которые исключает формат TimeML: *часто, время от времени*

временная последовательность

после, раньше, потом, затем (в виде формул для толкования *P* после *Q*, *P* раньше *Q*, *P*, *потом* / *затем Q*)

Примеры, которые исключает формат TimeML: *немедленно, мгновенно, сразу, предыдущий*

Несмотря на свою интуитивность, эти типы информации не являются семантически примитивными. Кроме того, между ними не существует четких границ. Каждый тип представляет собой область значений с ядром и обширной периферией, смыкающейся с соседними типами. Например, говоря о локализации, выделяют локализацию абсолютную и относительную. Абсолютная временная локализация означает, что положение на временной оси определено однозначно, например, *14 июля 2022 года*. Относительная локализация означает, что положение на временной оси задано относительно, например, *завтра* ‘на следующий день после дня, содержащего в себе момент речи’. Нередко бывает, что в качестве точки

отсчета берется другое событие, например, *ушел после обеда*. Однако это не только локализация, но и последовательность событий. Другая пара примеров:

(а) *приходил во вторник*

(б) *приходил по вторникам*

(а) — несомненная локализация, а (б) может быть классифицировано и как локализация повторяющегося события, и как периодичность. К рассмотрению этих примеров мы вернемся в четвертом разделе.

3. Описание формата

Разметка лексикализованных обозначений времени не сводится к выделению этих выражений в тексте. Существенно установить, как они функционируют, т.е. с какими словами в тексте связаны и каково значение (семантика) этих связей. Говоря о семантике связей, мы отдаем себе отчет, что строгое отделение ее от лексического значения единиц, образующих временное выражение, — задача трудная, может быть, нерешаемая. Однако некоторая нечеткость границ, на наш взгляд, не должна иметь серьезного влияния на формат: поскольку объект нашего исследования — *лексикализованные* обозначения времени, разметка с неизбежностью будет отражать какие-то элементы лексических значений рассматриваемых слов. Скорее, было бы странно, если бы при решении подобной задачи лексические значения релевантных слов полностью игнорировались.

Формат разметки предполагает выделение двух типов сущностей — собственно временных выражений и их связей. Учитывая опыт TimeML по созданию временной разметки “на голом тексте”, мы предлагаем выполнять ее “поверх” морфосинтаксической разметки. В качестве морфосинтаксической “основы” нами выбран формат корпуса СинТагРус [11], который синтаксически размечен в соответствии с моделью русского синтаксиса, разработанного в ИППИ РАН [12]. В техническом плане формат СинТагРус представляет собой xml, обогащенный специальными тегами и атрибутами, хотя пользователи обычно знакомятся со структурами СинТагРус в виде схем-графов, поскольку их проще воспринимать. Ниже такой граф приводится для примера (6). Примеры без указания источника взяты из СинТагРус. Примеры, источник которых указан в квадратных скобках, взяты из основного корпуса НКРЯ.

Предлагаемая временная разметка, как и TimeML, является расширением xml. Она

содержит всего два тега: TEMP и TEMP_RELATION. TEMP хранит информацию о временном выражении, TEMP_RELATION — о его связи с другими словами. Теги xml всегда образуют иерархию. Иерархически TEMP_RELATION является потомком TEMP. Иными словами, тег TEMP_RELATION вложен в тег TEMP. Содержательно это означает, что временное выражение, участвующее в связи, всегда фиксировано. Поэтому тег TEMP_RELATION имеет всего два атрибута: атрибут REL хранит информацию о типе связи, атрибут TARGET — информацию о втором участнике. Продемонстрируем разметку простых случаев из перечисленных выше.

(1) *28 сентября немцы заняли Варшаву.*

Внутри тега <S>, хранящего всю информацию о предложении, у этого примера имеется запись:

```
(1') <TEMP HEAD="1" WORDS="1,2"
ISO8601="XXXX-09-28">
  <TEMP_RELATION REL="localization"
  TARGET="56:4"/>
</TEMP>
```

У тега TEMP заполнено три атрибута: HEAD хранит информацию о синтаксической вершине временного выражения — это первое слово в предложении, WORDS содержит перечень номеров всех входящих в него слов, ISO8601 представляет координаты на временной оси в формате всемирного координированного времени (незаполненные позиции формата помечаются символом X). В целом выражение локализует во времени событие, называемое четвертым словом в этом же предложении, глаголом *заняли*. Этот факт отражен с помощью элемента <TEMP_RELATION> внутри элемента <TEMP>: наше выражение связано с четвертым словом в пятьдесят шестом предложении текста отношением типа localization. В значении атрибута TARGET, указывающего на второго участника связи, номеру слова (4) предшествует другая цифра — 56. Это номер данного предложения в тексте. Таким образом, формат предоставляет возможность устанавливать для временных выражений связи, выходящие за пределы предложения.

(2) *С августа 1942 года по февраль 1943 года японские и американские войска сражаются за контроль над островом Гуадалканал в составе архипелага Соломоновы острова.*

В данном предложении мы усматриваем два временных выражения — локализирующее начало события *сражаются* и локализирующее его конец.

```
(2') <TEMP HEAD="1" WORDS="1,2,3,4"
ISO8601="1942-08-XX">
  <TEMP_RELATION REL="start" TARGET="560:13"/>
</TEMP>
```



```
<TEMP HEAD="5" WORDS="5,6,7,8"
ISO8601="1943-02-XX">
<TEMP_RELATION REL="end" TARGET="560:13"/>
</TEMP>
```

Длительность и периодичность также представляются в виде специализированных отношений.

(3) *Тренеры думали **четыре часа**, но ничего не решили.*

Выражение *четыре часа* размечается следующим образом:

```
<TEMP HEAD="4" WORDS="3,4">
<TEMP_RELATION REL="duration"
TARGET="18:2"/>
</TEMP>
```

(4) *Врачи прописали мне **ежедневно** читать вслух "Евгения Онегина".*

```
<TEMP HEAD="4" WORDS="4">
<TEMP_RELATION REL="period" TARGET="6:5"/>
</TEMP>
```

Что касается последовательности событий во времени, то для нее в формате разметки предусмотрено не одно, а два отношения. События не существуют и не мыслятся иначе, как следующими друг за другом во времени. Поэтому стоит в речи или в тексте появиться нескольким обозначениям событий, как они тут же интерпретируются как последовательность: *пришел, увидел, победил*. Такие случаи мы не размечаем, так как наш объект исследования ограничен лексикализированными средствами выражения временной информации. Если последовательность обозначается лексически, то ее показатель связан не с одним, а с двумя событиями. Соответственно, для разметки используются два отношения, *sequence_1* и *sequence_2*. Первое отмечает более раннего участника, второе — более позднего.

(5) *Сливу отварите, **затем** протрите через сито.*

```
<TEMP HEAD="3" WORDS="3">
<TEMP_RELATION REL="sequence_1"
TARGET="7:2"/>
<TEMP_RELATION REL="sequence_2"
TARGET="7:4"/>
</TEMP>
```

4. Тонкости и трудности

Границы временных выражений

Мы уже говорили, что временная разметка выполняется "поверх" разметки синтаксической. Поэтому нам доступен простой и формальный способ определения границ временных выражений — временное выражение должно являться группой синтаксического узла. Под группой синтаксического узла понимается узел со всеми его синтаксическими зависимыми, непосредственными и опосредованными. Однако уже в примере (2)

мы видим, что границы поддереьев определяются не только семантикой, поэтому этот формальный способ годится не для всех случаев: выражения *с августа 1942 года* и *по февраль 1943 года* представляют группу узла *с*, так как предлог *по* зависит от него по соотносительному синтаксическому отношению. Семантически же эти два выражения независимы друг от друга. Каждое из них может функционировать в тексте само по себе, и смысл его при этом не меняется:

(2') *С августа 1942 года японские и американские войска сражаются за контроль над островом Гуадалканал в составе архипелага Соломоновы острова.*

(2'') *По февраль 1943 года японские и американские войска сражаются за контроль над островом Гуадалканал в составе архипелага Соломоновы острова. (однозначно возможно До февраля)*

Пример изолированного использования конструкции *с по* в рассматриваемом временном значении: *Создан мировой бренд несущей народам счастье "либеральной демократии", которая когда-то была и правда неплохой штукой (по И. Валлерстайну — по 1968 год), а сегодня стала очень красивой игрушкой внешне и своего рода разрушающим общества вирусом по сути.* [Михаил Рогожников. Какая демократия нам нужна (2004) // "Эксперт", 13.12.2004]

Рассмотрим еще один пример:

(6) *Но и спустя 100 лет после исследования Третьякова неизвестно, погибли ли клетки из-за скоплений вредного белка или у обоих процессов общая причина.*

Синтаксическая структура этого предложения следующая.

Начав рассмотрение с вершины *спустя*, мы увидим группу *и спустя 100 лет после исследования Третьякова*. Можно сказать, что это единое временное выражение, локализирующее состояние, обозначаемое словом *неизвестно*. Это будет корректным, но сильно огрубленным отражением функции этого выражения. Очевидно, что локализация в данном случае — относительная, основанная на точке отсчета, заданной как 'время, когда Третьяков выполнил свое исследование'. От этой точки отсчитывается сто лет, и на другом конце этого отрезка располагается локализуемое состояние.

Предлагаемый формат позволяет выполнить разметку этого предложения так, что оно будет содержать два временных выражения, одно из которых будет вложено в другое:

```
(6) <TEMP HEAD="3" WORDS="2,3,4,5,6,7,8">
</TEMP>
<TEMP HEAD="6" WORDS="6,7,8">
</TEMP>
```

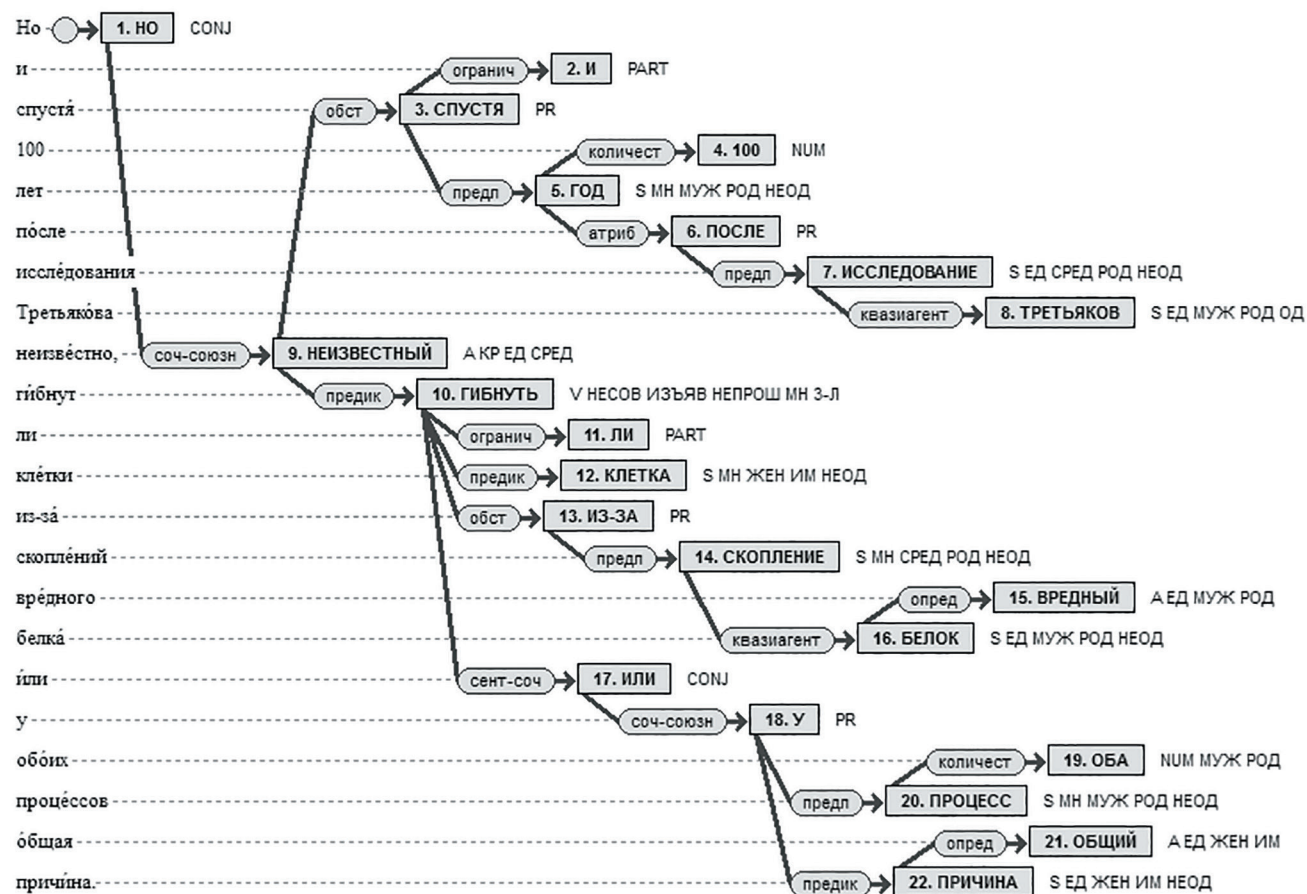



Рис. 1 Синтаксическая структура примера (6)

Полная разметка этого примера будет дана ниже, когда мы представим способ разметки разделяющих интервалов, а пока вернемся к рассмотрению границ временных выражений. Мы указали два возможных отклонения от типичного случая, когда временному выражению соответствует синтаксическая группа: случай, когда группа содержит два непересекающихся выражения, и случай, когда группа соответствует временному выражению, в которое “вложено” временное выражение меньшего размера. Но наибольшую трудность для описания представляют следующие сконструированные примеры:

(7а) Он *приедет* 5–6 июля.

(7б) *Фестиваль пройдет* 5–6 июля.

В (7а) выражение 5–6 июля локализует во времени событие приезда и содержит в себе дизъюнкцию: ‘приедет 5 июля или приедет 6 июля’. В (7б) то же самое выражение скорее фиксирует начальную и конечную точки: ‘фестиваль начнется 5 июля и закончится 6 июля’. Подобная интерпретация очевидно предпочтительна, когда даты не следуют друг за другом непосредственно: *Фестиваль пройдет*

5–12 июля. Синтаксически выражение 5–6 июля в обоих примерах устроено одинаково: вершиной является первое числительное, которое подчиняет себе название месяца по атрибутивному синтаксическому отношению, а второе числительное – по кратному синтаксическому отношению.

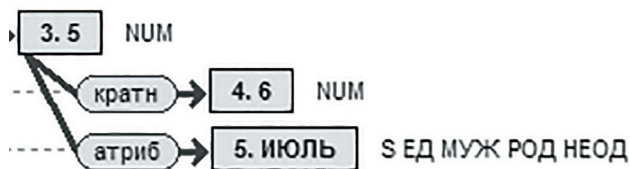


Рис. 2 Фрагмент синтаксической структуры примеров (7а) и (7б)

Поскольку тут имеется сочинительное сокращение (по смыслу второе числительное также связано с названием месяца), не хочется подходить к разметке примеров типа (7б) так же, как мы подходили к разметке примеров типа 2, содержащих информацию о начале и конце события: разделив поддерево на два непересекающихся выражения (5 июля и 6) таким образом, что первое обозначает начало, а второе – конец события, мы получим одно линейно разорванное выражение и одно неполное, которое не может самостоятельно

функционировать в предложении, передавая тот же смысл. То же самое относится к решению усматривать два выражения и в тех случаях, когда имеется дизъюнктивная интерпретация. Чтобы сохранить функциональность описания (под функциональностью мы понимаем то, что выделенное разметкой выражение является самодостаточной единицей — например, может быть перенесено в другое предложение с сохранением смысла), мы предлагаем в таких случаях видеть единое временное выражение. При этом все релевантные для него свойства, помимо границ в тексте, устанавливаются на семантических основаниях, а специфика соотношения морфосинтаксической структуры и семантики отражается с помощью специального атрибута COMPOSED. Таким образом, разметка выглядит так:

```
(76') <TEMP HEAD="3" WORDS="3,4,5"
ISO8601="XXXX-07-05" ISO8601="XXXX-07-06"
COMPOSED="YES">
  <TEMP_RELATION REL="start" TARGET="1:1"/>
  <TEMP_RELATION REL="end" TARGET="1:1"/>
</TEMP>
```

Разметка для (7a) выглядит так:

```
(7a') <TEMP HEAD="3" WORDS="3,4,5"
ISO8601="XXXX-07-05" ISO8601="XXXX-07-06"
COMPOSED="YES">
  <TEMP_RELATION REL="localization"
TARGET="1:2"/>
</TEMP>
```

Отношение дизъюнкции в разметке никак не отражается.

Выражения со значением интервала

Рассмотрим следующие примеры:

(8) Он пошел вниз и вернулся **через минуту** с парой великолепных защитных перчаток, предназначенных специально для фиксации кошек.

(9) Быстрая секретарша **на минуту** скрылась в кабинете и проговорила, вернувшись: — Товарищ Ефимова, пожалуйста.

В первом предложении минута — это отрезок времени между событиями, обозначенными глаголами *пошел* и *вернулся*, во втором — глаголом *скрылась* и деепричастием *вернувшись*. Если поставить во главу угла именно временной отрезок, то можно сказать, что пары *пошел* и *вернулся*, *скрылась* и *вернувшись* определяют границы интервала. Поэтому временные выражения в данных примерах получают следующую разметку:

```
(8') <TEMP HEAD="6" WORDS="6,7">
  <TEMP_RELATION REL="localization"
TARGET="1:5"/>
```

```
<TEMP_RELATION REL="interval after"
TARGET="1:2"/>
<TEMP_RELATION REL="interval before"
TARGET="1:5"/>
</TEMP>
```

```
(9') <TEMP HEAD="3" WORDS="3,4">
  <TEMP_RELATION REL="interval after"
TARGET="1:5"/>
  <TEMP_RELATION REL="interval before"
TARGET="1:10"/>
</TEMP>
```

Отношение *interval after* направлено к слову, которое задает левую границу интервала, его начало, а отношение *interval before* — к слову, которое задает его правую границу или конец.

Обращает на себя внимание своеобразная асимметрия: в примере (8) временное выражение синтаксически зависит от слова, ассоциированного с концом интервала, а в примере (9) — с началом. При этом естественно считать, что в примере (8) выражение также локализует во времени событие, завершающее интервал (*Когда он вернулся? — Через минуту*). В примере 9 нельзя сказать, что временное выражение локализует какое-то из связанных с ним событий: *Когда секретарша скрылась? *На минуту. Когда секретарша вернулась? *На минуту*.

Не всегда у интервалов охарактеризованы обе границы.

(10) И только **после 2014 года** оказалось достаточно политической воли у власти, чтобы это все закончить.

Здесь определена только левая граница интервала. Заметим, что это выражение также является локализацией, поскольку отвечает на вопрос *когда?* Таким образом, значение временного выражения в этом предложении следующее: 'интервал, ограниченный слева 2014 годом'.

```
(10') <TEMP HEAD="3" WORDS="3,4,5"
ISO8601="2014-XX-XX">
  <TEMP_RELATION REL="interval after"
TARGET="1:9:5"/>
  <TEMP_RELATION REL="localization"
TARGET="1:9:6"/>
</TEMP>
```

(11) В любом случае, сказала она, **до конца года** об этом не может быть и речи.

Определена только правая граница интервала. Временная разметка аналогична предыдущему примеру с точностью до имени отношения: в слову *конец* направлено отношение *interval before*.

Существенно, что сами границы в состав временных интервалов не входят: неверно из предложения (10) делать вывод, что у власти уже в 2014 году оказалось достаточно политической воли.

Вернемся теперь к примеру 6:

(6) *Но и спустя 100 лет после исследования Третьякова неизвестно, гибнут ли клетки из-за скоплений вредного белка или у обоих процессов общая причина.*

Временные выражения с предлогом *спустя* очень близки временным предложениям с предлогом *через*. Ср.

(6') *Но и через 100 лет после исследования Третьякова неизвестно, гибнут ли клетки из-за скоплений вредного белка или у обоих процессов общая причина.*

Соответственно, полная разметка объемлющего временного выражения *через 100 лет после исследования Третьякова* будет выглядеть так:

```
<TEMP HEAD="3" WORDS="2,3,4,5,6,7,8">
<TEMP_RELATION REL="localization"
TARGET="55:9"/>
<TEMP_RELATION REL="interval after"
TARGET="55:7"/>
<TEMP_RELATION REL="interval before"
TARGET="55:9"/>
</TEMP>
```

Аналогично примеру (8) отношение *interval after* направлено в слово *исследования*, так как именно оно обозначает событие, ассоциирующееся с границей интервала. Однако мы уже договорились, что это слово также входит в состав вложенного выражения *после исследования Третьякова*. Как соотносить это выражение с объемлющим его выражением? Мы предлагаем разметить вложенное выражение следующим образом:

```
<TEMP HEAD="6" WORDS="6,7,8">
<TEMP_RELATION REL="interval after"
TARGET="55:7"/>
</TEMP>
```

Содержательно это означает, что смысл выражения – ‘интервал, начатый исследованием Третьякова’

(в геометрических терминах – луч). Таким образом, это выражение размечается параллельно с временным выражением из примера (10). Кроме того, мы достигаем изоморфности временной разметки на синтаксическом и семантическом уровнях. Компонент `<TEMP_RELATION REL="interval after" TARGET="55:7"/>` повторяется в разметке объемлющего выражения, демонстрируя “вложенность” смысла меньшего выражения в большее.

Завершая разбор этого примера, обратим внимание читателя на то, что само по себе выражение *после исследования Третьякова* в другом контексте могло бы истолковываться как указание на событие в последовательности (*статья, опубликованная после исследования Третьякова*). Из указания на интервал всегда может быть выведена информация о временной последовательности событий. Таким образом примеры, размеченные с помощью *interval after* и *interval before*, оказываются на периферии выражений, передающих информацию о временной последовательности, на стыке этого типа с другими. Мы упоминали подобные периферийные явления, возникающие, когда одновременно передается временная информация разных типов, в разделе 2. Еще более ярко иллюстрирует связь между идеей разделяющего интервала и идеей последовательности наречие *сразу*. В своем временном значении это наречие обозначает отсутствие разделяющего интервала между двумя событиями. Ср. толкование из малого академического словаря русского языка [13] “В ближайший момент после чего-л.”. Многие контексты напоминают пример (6) – *сразу* сочетается с предлогом *после*.

(12) *Он вызвал отца и невесту телеграммой в Сибирь и сразу после венчания отбыл в Порт-Артур.*

Синтаксически это выражение сильно отличается от структуры выражения в (6): *сразу*

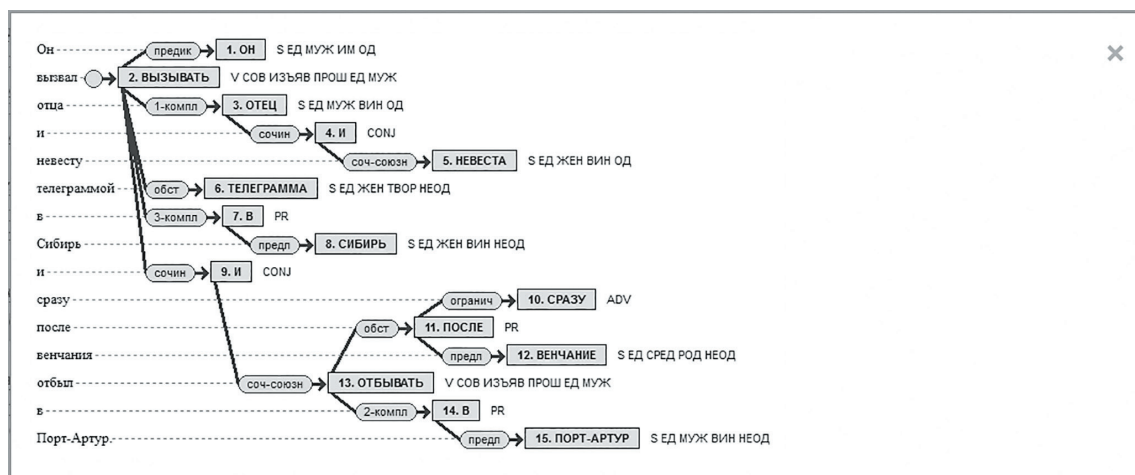


Рис. 3 Синтаксическая структура примера (12)

не управляет предлогом *после*, а зависит от него по ограничительному отношению.

Временная разметка должна, на наш взгляд, вскрывать общность семантических структур, скрывающихся за такими разными морфосинтаксическими образцами, как *сразу* + *после* и *через* + *после*. Применять обозначения *interval before* и *interval after* к описанию наречия *сразу* и ему подобных нам кажется неверным, поскольку они указывают на отсутствие интервала. Для этих случаев мы используем обозначения *lack of interval_1* и *lack of interval_2*.

```
(12') <TEMP HEAD="10" WORDS="10">
<TEMP_RELATION REL="lack of interval_1"
TARGET="49:12"/>
<TEMP_RELATION REL="lack of interval_2"
TARGET="49:13"/>
</TEMP>
<TEMP HEAD="11" WORDS="11,12">
<TEMP_RELATION REL="sequence_1"
TARGET="49:12"/>
<TEMP_RELATION REL="sequence_2"
TARGET="49:13"/>
</TEMP>
```

Пара *lack of interval_1* и *lack of interval_2* используется в разметке аналогично с парой *sequence_1* и *sequence_2*. Кроме того, на уровне логических выводов можно говорить о том, что события, к которым отсылают *lack of interval_1* и *lack of interval_2*, соотносятся так же, как и события, связанные отношениями *sequence_1* и *sequence_2*. В случае конструкции *сразу* + *после* этот факт в разметке отражается эксплицитно. Кроме того, то же логическое отношение существует между событиями или временными отрезками, связанными отношениями *interval before* и *interval after*, если эти отношения исходят из одного временно-го выражения.

Дейктические явления

Временная разметка не может игнорировать дейксис — механизм образования значения языковых средств за счет опоры на ситуацию речи. На данный момент мы ориентируем формат разметки на письменные тексты, однако с незначительными вариациями механизм дейксиса работает и в них. Так, в новостном тексте абсолютное значение слов *сегодня* и *завтра* выводится из знания о том, в какой день новость опубликовали. Подобную ситуацию можно назвать расширенным первичным дейксисом. С другой стороны, хорошо известно явление вторичного дейксиса, когда значение языковой единицы сходным образом опирается на сведения, имеющиеся в распоряжении понимающего, но отличные от

ситуации речи. Например, встречая те же самые слова *сегодня* и *завтра* в романе, читатель понимает, куда они помещают соответствующее событие — относительно некоторой точки вымышленного романного времени. Чтобы различить эти случаи, мы ввели в разметку два элемента, UT (Utterance Time, время, на которое ссылаются дейктические единицы в ситуации первичного дейксиса) и CTT (Current Text Time, альтернативное “якорное” время).

```
(76') <TEMP HEAD="3" WORDS="3,4,5"
ISO8601="X X X X-07-05" ISO8601="X X X X-07-06"
COMPOSED="YES">
<TEMP_RELATION REL="start" TARGET="1:1"/>
<TEMP_RELATION REL="end" TARGET="1:1"/>
</TEMP>
```

```
(7a') <TEMP HEAD="3" WORDS="3,4,5"
ISO8601="X X X X-07-05" ISO8601="X X X X-07-06"
COMPOSED="YES">
<TEMP_RELATION REL="localization"
TARGET="1:2"/>
</TEMP>
```

Эти специальные элементы дают нам возможность отразить вклад дейксиса в семантику выражения даже в тех случаях, когда у нас нет опоры для вычисления точного значения выражения (например, установить дату публикации разбираемой новости не представляется возможным). Теперь рассмотрим случай, когда нам известна дата, выступающая в качестве точки отсчета для употребленных в тексте дейктических средств, — она включена в конец текста или находится на месте подзаголовка. В этом случае разметку можно выполнять двумя способами: а) установить связь непосредственно с тем выражением, которое указывает на дейктическую точку отсчета; б) разметить дейктическое выражение так же, как в случае отсутствия точной информации о дейктическом центре, а референтному выражению присвоить атрибут, показывающий, что оно выступает в качестве дейктического центра. Мы предпочитаем второй способ, так как он позволяет единообразно размечать дейктические средства вне зависимости от знания о координатах дейктического центра и легко находить их при поиске. Соответствующий атрибут называется ANCHOR. В описанной выше ситуации разметка временно-го выражения, соответствующего дате публикации новости, содержит атрибут ANCHOR="UT". Кроме UT, этот атрибут может также принимать значение CTT. Выражения типа *однажды* и *как-то раз*, по нашему мнению, не столько локализуют во времени события, к которым относятся, сколько задают точку отсчета для последующего

изложения. Иногда эту функцию берет на себя целое предложение, например,

(13) *Шла Страстная неделя.*

Его мы размечаем следующим образом.

<TEMP HEAD="3" WORDS="2,3" ANCHOR="CTT">

Нелексистализованные события

Дейксис — базовый механизм языка: даже не зная конкретных значений точек отсчета, слушатель или читатель способен их сконструировать или реконструировать так, чтобы без потерь воспринять текст. Однако это лишь малая часть тех смыслов, которые не имеют конкретного лексического воплощения в тексте и которыми слушатель или читатель, тем не менее, свободно оперирует. Рассмотрим следующие примеры:

(14) *Я жду Васю завтра.*

(15) *Танцор на мгновение замер.*

(16) *Встреча отложена на неделю.*

(17) *Ветер метет поземку, мгновенно заравнивая лубой след.*

(14), (15), (16) и (17) включают в себя уже рассмотренные нами временные выражения или их квазисинонимы (*мгновенно* ≈ *сразу*). Их объединяет то, что на этапе установления семантической связи с событием возникает сложность.

Завтра в прототипическом случае локализует событие, обозначенное одним из глаголов предложения: сообщается, что оно произойдет на следующий день относительно дейктического центра. Однако в (14) это не так. *Завтра* локализует не событие, обозначенное глаголом *жду*, а появление Васи. В (15) *на мгновение* обозначает интервал. С глаголом *замер*, синтаксически управляющим временным выражением, соотносится начало этого интервала. Нетривиальное приращение смысла состоит в том, что концом интервала является прекращение состояния, в котором начал быть танцор. В (16) событие, обозначаемое глаголом *откладывать*, вообще никак не связано с границами и наполнением недельного временного интервала: неделя разделяет старую и новую даты встречи. В (17) есть указание на отсутствие интервала между заравниванием и появлением следа.

Эти примеры показывают, что, помимо уже упоминавшихся средств, инструменты временной разметки должны быть рассчитаны на работу с довольно широким спектром нелексистализованных смыслов. Для таких случаев наш формат

разрешает большую свободу при заполнении атрибута TARGET.

Это может быть фиктивное слово или короткое словосочетание с необходимым смыслом по усмотрению разметчика:

(14') <TEMP HEAD="4" WORDS="4" >
<TEMP_RELATION REL="localization" TARGET="приезд Васи"/>
<TEMP_RELATION REL="interval after" TARGET="CTT"/>
</TEMP>

(15') <TEMP HEAD="2" WORDS="2,3" >
<TEMP_RELATION REL="interval after" TARGET="62:4"/> (TARGET – *замер*)
<TEMP_RELATION REL="interval before" TARGET="завдвигался"/>
</TEMP>

Это может быть слово, уже связанное с выражением какой-то другой связью:

(16') <TEMP HEAD="3" WORDS="3,4" >
<TEMP_RELATION REL="interval after" TARGET="1:1"/> (TARGET – *Встреча*)
<TEMP_RELATION REL="interval before" TARGET="1:1"/> (TARGET – *Встреча*)
</TEMP>

Наконец, это может быть другой семантический компонент:

(17') <TEMP HEAD="4" WORDS="4" >
<TEMP_RELATION REL="lack of interval_1" TARGET="start"/>
<TEMP_RELATION REL="start" TARGET="84:7"/> (TARGET – *след*)
<TEMP_RELATION REL="lack of interval_2" TARGET="84:5"/> (TARGET – *заравнивая*)
</TEMP>

В разметку слова *мгновенно* включено отношение start, направленное в существительное *след*. За этой записью стоит такое рассуждение: значение существительного с предметным значением в конкретно-референтном употреблении всегда содержит утверждение 'существует X'. Существование, как любое другое состояние, может быть локализовано во времени, может иметь начало, конец, длительность. Соответственно получают временную разметку примеры типа *мода 30-х годов* 'мода, существовавшая в 30-е годы' (localization), *рукопись IX века* 'рукопись, созданная в IX веке' (start) и т.д. В рассматриваемом примере след существует и выражение *мгновенно* семантически относится именно к существованию следа: сообщается, что между его началом и концом нет никакого временного промежутка. А для представления этого смысла в нашем

формате есть стандартное средство – пара `lack of interval_1 / lack of interval_2`.

Временные выражения с порядковыми числительными

Рассмотрим примеры:

(18а) *На третий день по приезде мы узнали, что эти выстрелы, боевые возгласы и пролетевшее прямо над моей головой яйцо зовутся Хэллоуин.*

(18б) *Их батарея сделала первый залп только на третий день войны.*

(18в) *Так, на третий день я узнал о трагедии с бабабами.*

В них представлена конструкция с предлогом *на*, управляющим существительным в Вин. п. со значением единицы времени. У нее есть три варианта: а) от существительного зависит предложная группа с предлогом *по* или *после*; б) от существительного зависит существительное в родительном падеже; в) у существительного нет зависимых.

Во всех трех случаях временное выражение локализует во времени событие, от обозначения которого оно синтаксически зависит. Локализация эта относительная, она всегда опирается на точку отсчета. Эта ситуация содержательно совпадает с ситуацией, когда у нас имеется разделяющий интервал. В случае а) точка отсчета выражена эксплицитно, с помощью предложной группы.

Временная разметка примера:

```
(18а') <TEMP HEAD="1" WORDS="1,2,3,4,5" >
<TEMP_RELATION REL="localization"
TARGET="11:7"/>
<TEMP_RELATION REL="interval after"
TARGET="11:5"/>
<TEMP_RELATION REL="interval before"
TARGET="11:7"/>
</TEMP>
<TEMP HEAD="4" WORDS="4,5" >
<TEMP_RELATION REL="interval after"
TARGET="11:5"/>
</TEMP>
```

Разметка включает отношения *interval after* и *interval before*, тем самым показывая, что это выражение имеет общую семантическую часть с выражениями с предлогами *через* и *спустя* – оно точно так же опирается на разделяющий интервал.

В случае б) точкой отсчета является начало события, обозначаемого существительным в родительном падеже. В основе понимания – базовое знание: всякое событие обязательно имеет начало. Семантической компонент ‘начало’ вводится в разметку (18б) точно так же, как это было показано для (17):

```
(18б') <TEMP HEAD="7" WORDS="7,8,9,10" >
<TEMP_RELATION REL="localization"
TARGET="74:3"/>
<TEMP_RELATION REL="interval after"
TARGET="start"/>
<TEMP_RELATION REL="start" TARGET="74:10"/>
<TEMP_RELATION REL="interval before"
TARGET="74:3"/>
</TEMP>
```

В случае в) точка отсчета подразумевается – из текста “Маленького принца” читатель знает, что отсчет ведется от времени аварии. В разметке это время можно отразить, либо “вписав” появление в качестве значения атрибута (см. (14') и (15')), либо “направив” эту стрелку в одно из слов, обозначающих аварию, – в начале текста.

Временная разметка примера:

```
(18в') <TEMP HEAD="2" WORDS="2,3,4" >
<TEMP_RELATION REL="localization"
TARGET="227:6"/>
<TEMP_RELATION REL="interval before"
TARGET="227:6"/>
<TEMP_RELATION REL="interval after"
TARGET="авария"/>
</TEMP>
```

Рассмотрим теперь такой пример.

(19) *А об освобождении нас от имперского наваждения, заставляющего третье столетие подряд разрывать снарядами и бомбами клочок земли, населенный так и не покорившимся самым трудным для нас народом.*

Вершиной этого временного выражения не является предлог *на*, а семантически оно не является локализацией. Однако точно так же, как в примерах (18а–в), в основе его лежит точка отсчета и примыкающий к ней интервал некоторой длины. В нашем примере точкой отсчета является момент, когда ‘начали разрывать’, а длина интервала составляет больше 200, но меньше 300 лет. Однако, помимо неточности указания длины интервала, у этой конструкции имеется еще один семантический компонент. Чтобы его выявить, сравним конструкцию с выражениями *больше двухсот лет*, *больше двух столетий*. Они тоже обозначают интервал нечетко определенной длины. На наш взгляд, выражение *третье столетие подряд* отличается от выражений *больше двухсот лет*, *больше двух столетий подряд* по тому же признаку, по какому *давно* отличается от *долго*.

В статье Е.В. Падучевой, посвященной семантике наречий *давно* и *долго* [14], вводится такой различительный признак, как открытость/закрытость временного интервала. *Давно* – это обозначение большой длительности, у которого этот признак имеет значение “открытый”, а

долго — это обозначение большой длительности, у которого этот признак имеет значение “закрытый”. Открытость/закрытость интервала оценивается, исходя из законченности ситуации. Если ситуация не закончилась, то наблюдатель как бы наблюдает ее протекание или, по выражению Е.В. Падучевой, находится в синхронной позиции. Поэтому грамматически корректно *я давно тебя люблю*, и некорректно **я долго тебя люблю* — о незаконченности ситуации нам говорит форма непрошедшего времени несовершенного вида. А перевод этих утверждений в форму прошедшего времени меняет картину: ситуация становится законченной, и допустимость наречий *давно* и *долго* меняется на противоположную. Ср. *я долго тебя любила* (‘и перестала’) и несколько аномальное *я давно тебя любила*. Это предложение нуждается во временном контексте, задающем ретроспективную позицию наблюдателя, — например, *я уже тогда давно тебя любила*.

Протестируем в этих контекстах длительную конструкцию с порядковым числительным и длительную конструкцию с *больше*:

(20a) *Я третий год тебя люблю.*

(20б) *Я третий год тебя любила.*

(21a) *Я больше двух лет тебя люблю.*

(21б) *Я больше двух лет тебя любила.*

(20a) совершенно грамматично, (20б) требует расширения временного контекста, например, слова *тогда*. Иными словами, длительная конструкция с порядковым числительным в сочетании с формами глагола несовершенного вида ведет себя так же, как наречие *давно*. (21a) грамматично, (21б) грамматично и подразумевает ‘и перестала любить’. Таким образом, в отличие от наречия *долго*, длительная конструкция с *больше* оказывается приемлема в сочетании с формами несовершенного вида непрошедшего времени. Очевидно, она не охарактеризована по признаку открытости/закрытости интервала, поэтому может употребляться в контекстах обоих типов и соответственно пониматься. А вот интересующая нас конструкция с порядковым числительным по признаку открытости/закрытости интервала охарактеризована как обозначающая открытый интервал.

Мы приняли решение размечать открытые временные интервалы с помощью отношения *includes*, которое направлено либо в элемент *UT*, либо в элемент *СТТ*, поскольку сами по себе *UT* и *СТТ* можно рассматривать как обозначения позиции наблюдателя.

```
(19') TEMP HEAD="10" WORDS="9,10,11" >
<TEMP_RELATION REL="duration"
TARGET="136:12"/>
<TEMP_RELATION REL="includes"
TARGET="UT"/>
</TEMP>
```

Время наблюдения

Сравним два предложения:

(22a) *В девять часов Петя спал.*

(22б) *К девяти часам Петя спал.*

В чем состоит семантическое различие между ними?

(22a) означает:

в момент времени *t1*, называемый *девять часов*, имела место ситуация Р ‘Петя спит’, обозначаемая глаголом *спал*.

(22б) означает:

(а) в момент времени *t1*, называемый *девять часов*, имела место ситуация Р ‘Петя спит’, обозначаемая глаголом *спал*;

(б) ситуация Р начала иметь место в момент времени *t0*, предшествующий *t1*;

(в) интервал, разделяющий *t0* и *t1*, невелик.

Смысл ‘ситуация начала иметь место и продолжается’ — это смысл грамматической категории перфекта. Получается, что предложная группа, обозначающая время, не просто указывает время события, но и добавляет к смыслу предложения компонент (б), соответствующий грамматическому значению форм перфектного разряда. При этом нельзя сказать, что перфектом становится сама глагольная форма *спал*, так как от перфекта требуется обозначать более раннее событие относительно подразумеваемого позднего, а *спал* обозначает более позднее состояние относительно более раннего *заснул*, которое подразумевается выражением *к + N_{теп}*. Как каламбур можно сказать, что само выражение *к + N_{теп}* и есть такой неглагольный перфект.

Однако материалы, полученные в результате пилотной разметки корпуса, показывают, что выражение *к + N_{теп}* встречается не только при формах изъявительного наклонения несовершенного вида. Гораздо чаще оно сочетается с формами совершенного вида, финитными и нефинитными:

(23a) *Перегребли к вечеру с острова на материк [...].*

(23б) *Уже в марте 2006 года “Газета.Ру” со ссылкой на свои источники сообщила, что Кремль намерен выкупить “Коммерсант” у Патаркацишвили к парламентским выборам.*

(23в) *На уровне публикаций это не сказалось, и никто из читателей газеты не знал, что в 1998 году Яковлев, к тому времени серьезно увлекшийся буддизмом, покинул Россию, перебравшись в Лос-Анджелес.*

(23г) *К этому моменту на куске участка дачи Чуковского уже была построена детская библиотека, быстро ставшая центром светской и общественной жизни детской части переделкинского населения.*

Кроме того, это выражение может быть атрибутом существительного:

(24) *Еще раз для сравнения — на пресловутый проект “Роснефти” “Восток-Ойл”, который предполагает выход на уровень добычи в 100 млн тонн в год к 2030 году (по заверениям “Роснефти”), планируется потратить 10 трлн рублей и предоставить проекту 2,6 трлн рублей налоговых льгот.*

В этих контекстах ситуация, обозначаемая глагольной формой, соотносится не с моментом времени t_1 , а с моментом t_0 и подразумевает начало состояния, отличного от самой ситуации.

(23а) означает:

(а) в момент времени t_1 , называемый *вечер*, имела место ситуация Р ‘мы на материке’;

(б) ситуация Р начала иметь место в момент времени t_0 , предшествующий t_1 ;

(в) интервал, разделяющий t_0 и t_1 , невелик;

(г) ситуация Р начала иметь место в результате ситуации Q, обозначаемой глаголом *перегребли*;

(д) ситуация Q имела место в t_0 .

(24) означает:

(а) в момент времени t_1 , называемый *2030 год*, будет иметь место ситуация Р ‘уровень добычи составляет 100 млн тонн в год’;

(б) ситуация Р начнет иметь место в момент времени t_0 , предшествующий t_1 ;

(в) интервал, разделяющий t_0 и t_1 , невелик;

(г) смысл ‘Р начнет иметь место’ выражает существительное *выход*.

Обобщая толкования конкретных предложений, можно сказать, что $k + N_{temp}$ обозначает момент времени, когда имеет место результирующее состояние. При этом синтаксический хозяин предложной группы может обозначать как состояние, так и ситуацию, его вызвавшую.

Чтобы отличить случай, когда хозяин обозначает состояние, от случая, когда хозяин обозначает его начало/причину, можно использовать замену на стандартную локализирующую конструкцию. Для примера (22б) такой заменой будет конструкция с предлогом *в*, а результатом

замены — предложение (22а); для примера (23а) стандартной заменой будет слово *вечером*. В случае, когда синтаксический хозяин обозначает само состояние, замена на локализацию производит предложение, совместимое по смыслу с исходным: если к девяти часам Вася спал, то верно, что в 9 часов Вася спал. В противном случае такого соотношения установить нельзя: если к вечеру перегребли на материк, то нельзя с уверенностью сказать, что на материк перегребли вечером.

Случай, когда группа $k + N_{temp}$ обозначает время, когда наблюдается результирующее состояние, а ее синтаксический хозяин обозначает то, что это состояние начало, можно сказать, что выражение $k + N_{temp}$ локализует событие в интервале с известной правой границей, соответственно разметка будет выглядеть так (на примере (23а)):

```
(23а') <TEMP HEAD="2" WORDS="2,3">
<TEMP_RELATION REL="localization"
TARGET="312:1"/>
<TEMP_RELATION REL="interval before"
TARGET="312:3"/>
</TEMP>
```

Наличие в этом случае семантического компонента ‘локализация’ подтверждается тем фактом, что $k + N_{temp}$ в этом контексте может служить ответом на вопрос с *когда*: *Когда перегребли на материк? — К вечеру.*

Если синтаксический хозяин предложной группы обозначает само состояние, как в примере (22б), то говорить о локализации не приходится, ведь состояние не ‘вмещается’ в указываемый момент. Диагностика с помощью вопроса дает такой же результат: *Когда Петя спал? *К девяти часам.*

Мы предлагаем в таких случаях усматривать специальный тип временной информации, существующий на ряду с локализацией, длительностью, периодичностью и последовательностью — время наблюдения. На примере (22б) (*К девяти часам Петя спал*) продемонстрируем способ его обозначения:

```
(22б') <TEMP HEAD="1" WORDS="1,2,3"
ISO8601="XXXX-XX-XXT21:00:00">
<TEMP_RELATION REL="observation time"
TARGET="1:5"/>
<TEMP_RELATION REL="interval before"
TARGET="312:3"/>
</TEMP>
```

Кванторные временные наречия

Вслед за Е.В. Падучевой [15] мы признаем ряд наречий *никогда, редко, изредка, иногда, часто,*

вечно, все время, всегда имеющими специфические временные значения, которые позволяют говорить об этих наречиях как о кванторных временных наречиях. Приведем рассматриваемые ею примеры:

(25а) *Он всегда сидит в библиотеке.*

(25б) *Он всегда читает в автобусе.*

(25в) *Он всегда обедает в ресторане.*

Разная роль *всегда* в этих предложениях демонстрируется с помощью следующих перифраз:

(а) = 'Всегда, когда ни помотришь, он сидит в библиотеке'

(б) = 'Всегда, когда он находится в автобусе, он читает'

(в) = 'Всегда, когда он обедает, он делает это в ресторане'

На этом основании выделяются два временных значения, непрерывно-временное ('во все моменты времени', пример (25а)) и дискретно-временное ('каждый раз, когда', примеры (25б) и (25в)). *Всегда* в первом значении Е.В. Падучева также характеризует как квантор общности по моментам времени, а во втором — как квантор по событиям.

Исследование того, какие предикаты выступают в качестве основания для квантификации и какими грамматическими средствами они оформляются, представляется нам самостоятельной интересной задачей, поэтому в схему разметки мы включили средства, позволяющие отразить соответствующие отношения. Формат предусматривает два типа связей, *quantifier_1* и *quantifier_2*. *quantifier_1* связывает наречие с основанием для квантификации, *quantifier_2* — с квантуемым предикатом. Соответственно, в предложении (в) отношение *quantifier_1* связывает слово *всегда* с глаголом *обедает*, а отношение *quantifier_2* — с предлогом *в*, синтаксической вершиной обстоятельства *в ресторане*.

(25в') <TEMP HEAD="2" WORDS="2" >
<TEMP_RELATION REL="quantifier_1"
TARGET=":3"/>
<TEMP_RELATION REL="quantifier_2"
TARGET=":4"/>
</TEMP>

В предложении (б) отношение *quantifier_1* связывает слово *всегда* с предлогом *в*, синтаксической вершиной обстоятельства *в автобусе*, а отношение *quantifier_2* — с глаголом *читает*.

(25б') <TEMP HEAD="2" WORDS="2" >
<TEMP_RELATION REL="quantifier_1"
TARGET=":4"/>

<TEMP_RELATION REL="quantifier_2"
TARGET=":3"/>
</TEMP>

Вхождения *всегда* в непрерывно-временном значении мы предлагаем размечать с помощью единственной связи, *quantifier_2*. Таким образом, отсутствие связи *quantifier_1* при наличии *quantifier_2* следует понимать в том смысле, что по умолчанию имеет место квантификация по моментам времени. Соответственно, в примере (а) отношение *quantifier_2* связывает наречие *всегда* с глаголом *сидит*.

(25а') <TEMP HEAD="2" WORDS="2" >
<TEMP_RELATION REL="quantifier_2"
TARGET=":3"/>
</TEMP>

Заключение

Объем пилотной разметки, выполненной в соответствии с предлагаемым форматом, составляет всего 28 тысяч слов. Это 1611 предложений, из которых 652 содержат временную разметку. Всего размечено 966 временных выражений. Этого количества совершенно недостаточно, чтобы делать статистические обобщения, однако мы предполагаем, что в разметке представлены все основные морфосинтаксические конструкции с временной семантикой. Естественным следующим шагом является выработка морфосинтаксических шаблонов, которые можно отыскать в СинТагРусе автоматическими средствами, выполнение таким образом черновой темпоральной разметки во всем корпусе и экспертная ее проверка. Корпус объемом 1,5 миллиона слов, снабженный темпоральной разметкой, откроет перед лингвистами возможности для количественного описания функционирования временных выражений в тексте, а также для исследования количественных закономерностей их сочетания с различными грамматическими категориями, что, как мы надеемся, послужит вкладом в русскую аспектологию.

СПИСОК ЛИТЕРАТУРЫ

1. *The Proposition Bank (PropBank)* <http://propbank.github.io>
2. Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A., Webber B. The Penn Discourse TreeBank 2.0. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).

3. Grishman R., Sundheim B. Message Understanding Conference – 6: A Brief History. Proceedings of the 16th International Conference on Computational Linguistics (COLING), I, Copenhagen, 1996.
4. Translingual Information Detection, Extraction, and Summarization (TIDES) Evaluation Site. <https://www-nlpir.nist.gov/tides/index.html>
5. Doddington G.R., Mitchell A., Przybocki M.A., Ramshaw L.A., Strassel S.M., Weischedel R.M. The automatic content extraction (ace) program-tasks, data, and evaluation. LREC 2004 (Vol. 2, No. 1).
6. Ferro L., Gerber L., Mani I., Sundheim B., Wilson G. TIDES. Instruction Manual for the Annotation of Temporal Expressions. Technical Report Interim Draft for Terqas Workshop, The MITRE Corporation, 2002.
7. Cavar D., Dickson B., Aljubailan A., Kim S. Temporal Information and Event Markup Language: TIE-ML Markup Process and Schema Version 1.0. arXiv preprint arXiv:2109.13892 (2021).
8. Reichenbach H. Elements of symbolic logic. N.Y.: The MacMillan Co, 1947.
9. Sauri R., Littman J., Knippen B., Gaizauskas R., Setzer A., Pustejovsky J. TimeML Annotation Guidelines Version 1.2. 1., 2006.
10. Апресян Ю.Д. К новой версии теории лексических функций (ЛФ) // Международная конференция, посвященная 50-летию Петербургской типологической школы: Материалы и тезисы докладов. СПб., 2011.
11. Иншакова Е.С., Иомдин Л.Л., Митюшин Л.Г., Сизов В.Г., Фролова Т.И., Цинман Л.Л. СинТагРус сегодня // Труды Института русского языка им. В.В. Виноградова (21). М., 2019. С. 14–41.
12. Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л., Санников В.З. Теоретические проблемы русского синтаксиса: Взаимодействие грамматики и словаря. М.: Языки славянских культур, 2010.
13. Словарь русского языка: в 4 т. 2-е изд., испр. и доп. М.: Русский язык, 1981–1984.
14. Падучева Е.В. Давно и долго // Логический анализ языка: язык и время. М.: Индрик, 1997.
15. Падучева Е.В. Высказывание и его соотносительность с действительностью: Референциальные аспекты семантики местоимений. Изд. 6-е, испр. М.: Издательство ЛКИ, 2010.
- TreeBank 2.0. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).
3. Grishman, R., Sundheim, B. *Message Understanding Conference – 6: A Brief History*. Proceedings of the 16th International Conference on Computational Linguistics (COLING), I, Copenhagen, 1996.
4. *Translingual Information Detection, Extraction, and Summarization (TIDES) Evaluation Site*. <https://www-nlpir.nist.gov/tides/index.html>
5. Doddington, G.R., Mitchell, A., Przybocki, M.A., Ramshaw, L.A., Strassel, S.M., Weischedel, R.M. *The automatic content extraction (ace) program-tasks, data, and evaluation*. Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC 2004 (Vol. 2, No. 1).
6. Ferro, L., Gerber, L., Mani, I., Sundheim, B., Wilson, G. *TIDES. Instruction Manual for the Annotation of Temporal Expressions*. Technical Report Interim Draft for Terqas Workshop, The MITRE Corporation, 2002.
7. Cavar, D., Dickson, B., Aljubailan, A., Kim, S. Temporal Information and Event Markup Language: TIE-ML Markup Process and Schema Version 1.0. arXiv preprint arXiv:2109.13892 (2021).
8. Reichenbach, H. *Elements of symbolic logic*. N.Y.: The MacMillan Co, 1947.
9. Sauri, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., Pustejovsky, J. TimeML Annotation Guidelines Version 1.2. 1., 2006.
10. Apresyan, Ju.D. *Knovoj versii teorii leksicheskikh funkcij (LF)* [Preliminary Remarks about a New Version of the Theory of Lexical Functions (LF)]. *Mezhdunarodnaja konferencija, posvjashchennaja 50-letiju Peterburgskoj tipologicheskoj shkoly: Materialy i tezisy dokladov*. [International Conference Dedicated to the 50th Anniversary of the St. Petersburg Typological School: Materials and Abstracts]. St. Petersburg, 2011. (In Russ.)
11. Inshakova, E.S., Iomdin, L.L., Mityushin, L.G., Sizov, V.G., Frolova, T.I., Cinman, L.L. *SinTagRus segodnja* [The SynTagRus Today]. *Trudy Instituta russkogo jazyka im. V.V. Vinogradova* [Proceedings of the V.V. Vinogradov Institute of the Russian Language]. 2019, Iss. 21, pp. 14–41. (In Russ.)
12. Apresyan, Yu.D., Boguslavsky, I.M., Iomdin, L.L., Sannikov, V.Z. *Teoreticheskie problemy russkogo sintaksisa: Vzaimodeistvie grammatiki i slovary* [Theoretical Issues of Russian Syntax: Grammar–Lexicon Interaction]. Moscow, Yazyki Slavyanskikh Kultur Publ., 2010. (In Russ.)
13. *Slovar russkogo jazyka* [The Dictionary of Russian], 2nd ed. Moscow, Russkij jazyk Publ., 1981–1984. (In Russ.)

REFERENCES

1. *The Proposition Bank (PropBank)* <http://probank.github.io>
2. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, B. *The Penn Discourse*

14. Paducheva, E.V. *Davno i dolgo* [*Davno and Dolgo* (rus. 'a long time')]. *Logicheskij analiz jazyka: jazyk i vremja* [Logical Analysis of Language: Language and Time]. Moscow, Indrik Publ., 1997. (In Russ.)
15. Paducheva, E.V. *Vyskazyvanie i ego sootnesennost s dejstvitelnostju: Referencialnye aspekty semantiki mestoimenij* [Sentence in Its Relation to Reality: Referential Aspects of the Pronominal Semantics]. Moscow, Izdatelstvo LKI Publ., 2010. (In Russ.)

Дата поступления материала в редакцию: 20 сентября 2022 г.
Статья поступила после рецензирования и доработки: 8 октября 2022 г.
Статья принята к публикации: 15 октября 2022 г.
Дата публикации: 31 декабря 2022 г.

Received by Editor on September 20, 2022
Revised on October 8, 2022
Accepted on October 15, 2022
Date of publication: December 31, 2022