I. РАЗВИТИЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА И ДРУГИХ КОРПУСОВ

[1]Е.С. Иншакова, [2]Л.Л. Иомдин, [3]Л.Г. Митюшин,
[4]В.Г. Сизов, [5]Т.И. Фролова, [6]Л.Л. Цинман
[123456]Институт проблем передачи информации им. А. А. Харкевича РАН,
[2]Российский государственный гуманитарный университет
(Россия, Москва)
[1]e.s.inshakova@gmail.com, [2]iomdin@gmail.com, [3]mit@iitp.ru,
[4]sizov@iitp.ru, [5]tfrolova@gmail.com, [6]llcinman@gmail.com

СИНТАГРУС СЕГОДНЯ*

В статье описывается современное состояние корпуса СинТагРус, содержащего русские тексты с морфосинтаксической разметкой. На различных стадиях работы над корпусом были введены дополнительные типы разметки: лексико-семантическая, лексико-функциональная, анафорическая и микросинтаксическая.

Данные морфосинтаксической разметки предложения включают морфологический разбор каждого слова и синтаксическую структуру предложения в виде дерева зависимостей в соответствии с моделью «Смысл ↔ Текст» И. А. Мельчука и А. К. Жолковского. Лексико-семантическая разметка предполагает, что для каждого слова указана соответствующая ему статья комбинаторного словаря русского языка. Лексико-функциональная разметка представляет собой выделение в текстах словосочетаний, допускающих интерпретацию в терминах лексических функций. Результатом анафорической разметки является маркирование антецедентов местоимений. Микросинтаксическая разметка идентифицирует встретившиеся в текстах синтаксические фраземы и некоторые нестандартные синтаксические конструкции.

Разметка новых текстов корпуса выполняется в несколько стадий. Вначале тексты обрабатываются многофункциональным лингвистическим процессором ЭТАП-4, который в автоматическом режиме вносит в них морфосинтаксическую и лексико-семантическую разметку. Затем результаты работы процессора проверяются и при необходимости корректируются квалифицированными лингвистами-аннотаторами. После этого ЭТАП-4, используя построенные морфосинтаксические структуры, выполняет лексико-функциональную и анафорическую разметку.

* Данная работа поддержана грантом РФФИ № 19-07-00842 и Программой Президиума РАН «Памятники материальной и духовной культуры в современной информационной среде».

Лингвисты проверяют и корректируют эти типы разметки и вручную вносят в тексты микросинтаксическую разметку.

Корпус СинТагРус используется для целей теоретико-лингвистических исследований, а также для практической лексикографии. Статистика корпуса может учитываться при автоматической обработке текстов для оптимизации принимаемых решений. Весьма перспективно использование данных корпуса в системах, основанных на машинном обучении.

Ключевые слова: СинТагРус, синтаксически размеченный корпус, корпус русских текстов, грамматика зависимостей, лексические функции, антецеденты местоимений, микросинтаксис, эллипсис.

1. Общие сведения

Глубоко аннотированный корпус русских текстов СинТагРус разрабатывается в Лаборатории компьютерной лингвистики ИППИ РАН им. А. А. Харкевича начиная с 1998 г. При подготовке корпуса используется лингвистический процессор ЭТАП-4 (подробнее о нем см. [Apresjan et al. 2003]). Доступ к корпусу имеется на сайте русского национального корпуса (НКРЯ). На сайте доступны морфологическая (со снятой омонимией), синтаксическая (в терминах деревьев зависимостей) и лексико-функциональная разметка. Кроме того, для корпуса разработаны другие типы разметки: лексическая (или лексико-семантическая — разрешение лексической неоднозначности), анафорическая и разметка микросинтаксических конструкций. Эти типы разметки пока не доступны онлайн.

В состав корпуса входит более 650 художественных, публицистических, научно-популярных и новостных текстов. Разговорная речь, поэтические и технические тексты в корпусе не представлены, что позволяет добиться некоторой однородности представленного материала. На данный момент (конец 2019 г.) в корпусе содержится более 1,100,000 словоупотреблений, около 77 тысяч предложений. Корпус постоянно пополняется.

СинТагРус является единственным в мире полностью отредактированным экспертами-лингвистами корпусом текстов на русском языке с аннотацией на морфосинтаксическом уровне.

Этапы развития корпуса отражены в работах [Boguslavsky et al. 2000; Апресян и др. 2005; Богуславский и др. 2008(а,б); Шеманаева, Фролова 2010; Boguslavsky 2014; Дяченко и др. 2015].

Ниже перечислены и кратко рассмотрены виды разметки, уже описанные в предыдущих работах: морфологическая (раздел 2.1), синтаксическая, в том числе разметка эллиптических конструкций (разделы 2.2 и 2.3), лексико-семантическая (раздел 2.4), лексико-функциональная (раздел 2.5). Затем более подробно рассмотрены два новых типа разметки: анафорическая разметка (раздел 3.1) и разметка микросинтаксических конструкций (раздел 3.2). Кроме того, приведены сведения об использовании корпуса СинТагРус для решения исследовательских и прикладных задач (раздел 4).

2. Традиционные типы разметки СинТагРус'а

Разметка текстов в СинТагРус'е производится следующим образом: для подготовки к разметке текст обрабатывается программой сегментации текста, которая автоматически разбивает его на отдельные предложения; после этого каждое предложение обрабатывается морфологическим и синтаксическим анализатором многофункционального лингвистического процессора ЭТАП-4. В результате этой обработки формируется морфологическая структура для каждого слововхождения и синтаксическая структура для каждого предложения (см. разделы 2.1 и 2.2). Одновременно выполняется разрешение лексической омонимии (см. раздел 2.4). После автоматического этапа обработки все предложения проверяются лингвистом-аннотатором, который вносит необходимые изменения в структуры слов и предложений. При редактировании в предложения при необходимости вручную вносятся сведения об эллиптических конструкциях (см. раздел 2.3). Такая проверка, хотя и трудоемкая, позволяет достичь высокого уровня точности разметки. При корректировке текста постоянно вносятся уточнения в правила и словари ЭТАП-4. Лексико-функциональная (см. раздел 2.5) и анафорическая (см. раздел 3.1) разметка, а также разметка микросинтаксических конструкций (см. раздел 3.2) выполняются для предложений с построенной и откорректированной синтаксической структурой, при этом лексико-функциональная и анафорическая разметка выполняются в автоматическом режиме с последующей ручной коррекцией, а разметка микросинтаксическими конструкциями производится полностью вручную.

2.1. Морфосинтаксическая и лексическая разметка

При морфологической разметке каждое слововхождение снабжается сведениями о его морфологической структуре. Морфологическая структура словоформы представляет собой имя лексемы вместе с информацией о части речи и списком морфологических характеристик. Так, структура словоформы *вытесняя* имеет следующий вид: ВЫТЕСНЯТЬ, V НЕСОВ ДЕЕПР НЕПРОШ. Здесь ВЫТЕСНЯТЬ — имя лексемы, V обозначает глагол, НЕСОВ — несовершенный вид, ДЕЕПР — деепричастие, НЕПРОШ — непрошедшее (настоящее-будущее) время. Полный список частей речи, русских морфологических категорий и характеристик приводится на сайте НКРЯ в разделе «Инструкция к синтаксически размеченному корпусу».

2.2. Синтаксическая разметка

Результатом синтаксической разметки является дерево зависимостей, в узлах которого стоят слова предложения, а ветви помечены именами синтаксических отношений. Такое представление о синтаксической структуре предложения восходит к лингвистической модели «Смысл ↔ Текст» И. А. Мельчука

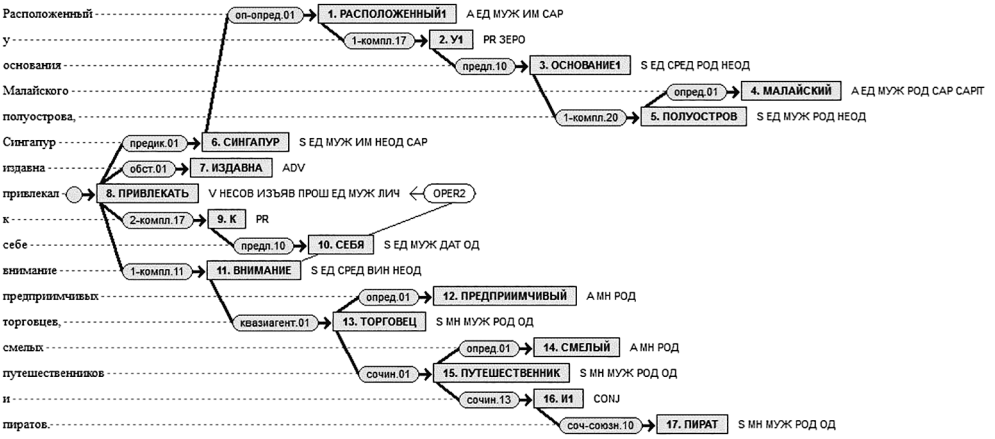

Рис. 1. Синтаксическая структура предложения (1)

и А. К. Жолковского [Жолковский, Мельчук 1967; Мельчук 1974]. В корпусе различается около 70 типов синтаксических отношений. Такая подробная разметка позволяет сделать описание синтаксической структуры более полным и лингвистически содержательным, чем в большинстве существующих корпусов. Перечень синтаксических отношений представлен в онлайновой версии СинТагРус'а на сайте НКРЯ в том же разделе, что и морфологические сведения.

Рассмотрим синтаксическую структуру следующего предложения:

(1) *Расположенный у основания Малайского полуострова, Сингапур издавна привлекал к себе внимание предприимчивых торговцев, смелых путешественников и пиратов*.

В узлах этого дерева зависимостей стоят слова предложения, представленные именами лексем (в прямоугольниках) и цепочками грамматических характеристик (справа от прямоугольников), а ветви помечены именами синтаксических отношений (в овалах).

Номера при именах синтаксических отношений указывают на номера правил в блоке синтаксических правил процессора ЭТАП-4, которые установили эти отношения. Как правило, эти номера соответствуют разным типам конструкций, в которых такое отношение может быть установлено. Так, сочинительное отношение, соответствующее бессоюзному сочинению однородных членов предложения, и сочинительное отношение, соответствующее сочинению однородных членов с помощью союза И, устанавливаются разными правилами. Первый тип представлен в структуре предложения (1) отношением сочин.01, которое соединяет узлы 13 (*торговцев*) и 15 (*путешественников*). Второй тип можно видеть в том же предложении между узлами 15 (*путешественников*) и 16 (*и*), это отношение обозначено как сочин.13. Эта информация не представлена в онлайн-версии корпуса.

Номера при некоторых именах лексем являются результатом лексико-семантической разметки. О ней см. ниже в разделе 2.4.

Стрелка с овалом, содержащим запись OPER2, соответствует лексико-функциональному отношению, устанавливаемому между узлами 11 и 8. О лексико-функциональной разметке см. ниже в разделе 2.5.

Вершиной предложения (1) является глагол ПРИВЛЕКАТЬ (в форме мужского рода, единственного числа, прошедшего времени, изъявительного наклонения, несовершенного вида — вид трактуется в лингвистическом процессоре ЭТАП-4 как словоизменительная характеристика). У вершинного глагола заполнены три синтаксические валентности. Предикативное синтаксическое отношение (на рисунке «предик.01») связывает глагол с подлежащим СИНГАПУР (неодушевленное существительное мужского рода в форме единственного числа, именительного падежа). Первое комплетивное отношение (1-компл.11) связывает вершинный глагол со вторым актантом — существительным ВНИМАНИЕ (неодушевленное существительное среднего рода в форме единственного числа, винительного падежа). Второе комплетивное отношение (2-компл.17) связывает глагол с вершиной предложной группы (предлог К), являющейся третьим актантом. От предлога К идет предложное отношение (предл.10) к существительному СЕБЯ (в ЭТАП-4 для него принята трактовка как одушевленного существительного мужского рода, в данном случае в единственном числе, дательном падеже). Описательно-определительное отношение связывает подлежащее СИНГАПУР с обособленным определением «расположенный у основания Малайского полуострова», в свою очередь имеющим внутреннюю синтаксическую структуру, где «у основания» является дополнением (1-компл.17) прилагательного РАСПОЛОЖЕННЫЙ, а «полуострова» дополнением (1-компл.20) существительного ОСНОВАНИЕ1. Квазиагентивное отношение (квазиагент.01), предназначенное для указания на субъект предикатного слова, не являющегося глаголом, связывает существительное ВНИМАНИЕ с вершиной сочинительной группы *предприимчивых торговцев, смелых путешественников и пиратов*, внутри которой однородные члены предложения связаны сочинительным отношением.

2.3. Особые случаи синтаксической разметки. Разметка предложений с эллипсисом

Чуть более 2%, почти полторы тысячи, предложений в СинТагРус'е описываются в синтаксической структуре как содержащие эллипсис. В этих случаях аннотатором принято решение о том, что для адекватного описания синтаксической структуры предложения в нее необходимо внести узлы, не соответствующие никаким словам в тексте предложения. Такие узлы называются «фантомными».

Чаще всего имена лексем для восстановленных узлов совпадают с именами лексем узлов, уже встречавшихся в предложении. Рассмотрим предложение (2).

(2) *По данным Всемирной организации здравоохранения, впервые в истории современной России средняя продолжительность жизни сильного пола достигла 66 лет, а у представительниц прекрасного пола — 77 лет*.

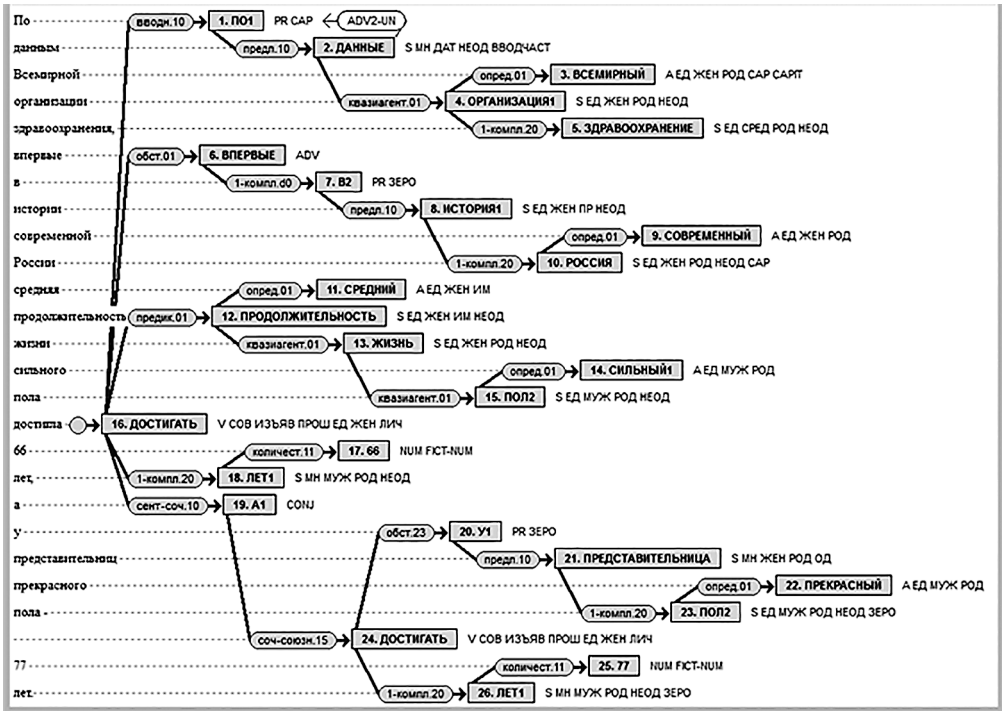

Рис. 2. Синтаксическая структура предложения (2)

В этом сложносочиненном предложении сентенциально-сочинительное отношение (сент-соч.10) связывает вершинный глагол (узел 16) ДОСТИГАТЬ в форме женского рода, единственного числа, прошедшего времени, изъявительного наклонения, совершенного вида с сочинительным союзом А (узел 19). А сочинительно-союзное отношение (соч-союзн.15) связывает этот союз А с фантомным узлом 24, у которого морфологическая структура совпадает с морфологической структурой вершинного узла.

В некоторых случаях восстанавливаются глагольные узлы с размытой семантикой — так называемые неопределенные глаголы. В этих случаях в качестве имени лексемы пишется НЕОПР-ГЛАГОЛ, а в скобках предлагается вариант, представляющийся уместной гипотезой. Это интересное, но не очень распространенное явление — таких узлов 206 во всем СинТагРус'е. Рассмотрим предложение (3) и его структуру.

(3) *Остановились на таком маршруте: электричкой до Серпухова (в абсолютно переполненном вагоне), там от вокзала автобусом до пристани, потом на пароходе до Велегожа и оттуда тропкой (4 км) по прекрасному лесу вдоль Оки до заветной поляны.*

В этом предложении восстановлены четыре фантомных глагольных узла с примерным смыслом 'добираться' — 6, 16, 23, 28, все они связаны в сочинительную цепочку.

Более подробно представление эллиптических конструкций рассмотрено в [Дяченко и др. 2015].

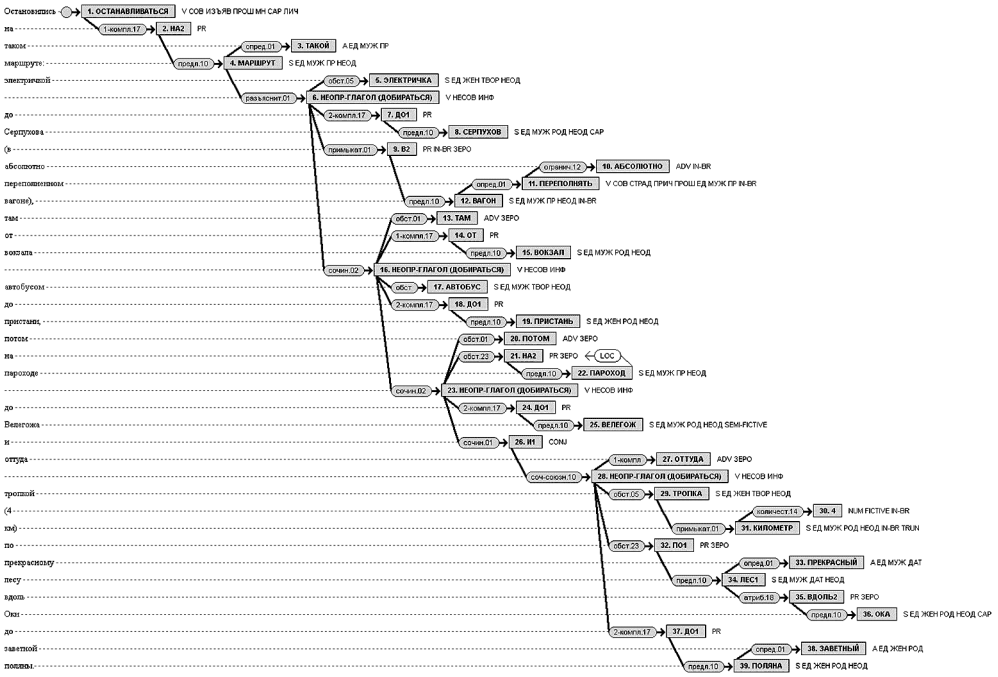

Рис. 3. Синтаксическая структура предложения (3)

2.4. Лексико-семантическая разметка и разрешение лексической неоднозначности

При лексико-семантической разметке для каждого многозначного слова в предложении выбирается одно из значений, имеющихся в комбинаторном словаре русского языка процессора ЭТАП-4. Номера при лексемах отсылают к разным статьям этого комбинаторного словаря.

Так, в предложении (1) индекс 1 в РАСПОЛОЖЕННЫЙ1 означает, что парсер (или лингвист-аннотатор при редактировании) выбрал первое значение слова РАСПОЛОЖЕННЫЙ ‘находящийся’, а не значение ‘склонный’ (ср. *расположенные к веселью дети*). Индекс 1 в У1 означает, что выбран предлог У, а не омонимичное междометие (ср. *У, как страшно!*). Индекс 1 в ОСНОВАНИЕ1 означает, что выбрано первое значение существительного ОСНОВАНИЕ ‘фундамент, нижняя опорная часть’, а не второе ‘момент возникновения, начало существования’ (ср. *Основание университета послужило толчком к развитию города*). Индекс 1 в И1 означает, что выбран союз И, а не омонимичная ему частица (ср. *Так и случилось*).

В предложении (2) в результате лексико-семантической разметки была разрешена лексическая омонимия в следующих случаях. Выбран омоним слова ОРГАНИЗАЦИЯ со значением ‘учреждение, предприятие’, а не со значением ‘процесс формирования’ (ср. *организация поставок продовольствия*). Выбран

омоним предлога В, управляющий предложным падежом (ср. *всё дело в лесе*), а не управляющий винительным падежом (ср. *идти в лес*) и не управляющий местным падежом (ср. *гулять в лесу*). Выбран омоним полнозначного прилагательного СИЛЬНЫЙ, а не часть сложного слова (ср. *200-сильный мотор*). Для слова ПОЛ выбран омоним ‘один из двух генетически противопоставленных разрядов живых существ’, а не ‘нижняя часть помещения’ (ср. *паркетный пол*) и не ‘половина’ (ср. *полгруши*), а также не ‘мужское имя’ (ср. *Пол Хьюитт написал книгу*). Для слова ЛЕТ выбран омоним ‘годов’ (родительный падеж, множественное число), а не отглагольное существительное со значением ‘полет’ (ср. *Снежинки таяли на лету*). Для слова А выбран союз, а не омонимичная частица (ср. *А кто это сказал?*). Для слова У выбор омонима такой же, как в предложении (1).

Информация о разрешении лексической омонимии пока недоступна в онлайновой версии СинТагРус’а.

2.5. Лексико-функциональная разметка

При лексико-функциональной разметке обнаруживаются и отмечаются в тексте словосочетания, допускающие интерпретацию в терминах лексических функций.

Под лексическими функциями (ЛФ) в соответствии с теорией ЛФ И. А. Мельчука, А. К. Жолковского и Ю. Д. Апресяна понимаются смыслы, для которых способ выражения определяется не только самим этим смыслом, но и словом-аргументом, при котором выражается этот смысл. Так, значение ЛФ MAGN ‘высокая степень или интенсивность’ выражается в русском языке прилагательным ПРОЛИВНОЙ при существительном ДОЖДЬ и прилагательным ВОЛЧИЙ при существительном ГОЛОД. Выше в предложении (1) значением ЛФ OPER2 при существительном ВНИМАНИЕ является глагол ПРИВЛЕКАТЬ, в то же время для существительного ИНТЕРЕС значением этой же ЛФ является глагол ПРЕДСТАВЛЯТЬ. Аппарат ЛФ удобно использовать в автоматическом переводе, поскольку ЛФ кодируют универсальные смыслы, по-разному представленные в разных языках и при разных аргументах. Корпус, снабженный ЛФ-разметкой, позволяет исследовать контексты, в которых реализуются ЛФ.

В корпусе представлено 129 различных ЛФ. Полный список ЛФ с пояснениями и примерами можно видеть на сайте НКРЯ. На 2019 г. в корпусе отмечено свыше 21 тысячи предложений, содержащих около 30 тысяч ЛФ-сочетаний.

ЛФ-разметка, так же как и морфосинтаксическая и лексическая разметка, производится в автоматическом режиме на построенных и отредактированных синтаксических структурах: вначале ЛФ-анализатор лингвистического процессора ЭТАП-4 на основании сведений, записанных в комбинаторном словаре и в правилах распознавания ЛФ-сочетаний, выделяет ЛФ-сочетания в предложении. Затем полученный результат проверяется и при необходимости исправляется и дополняется лингвистом-аннотатором.

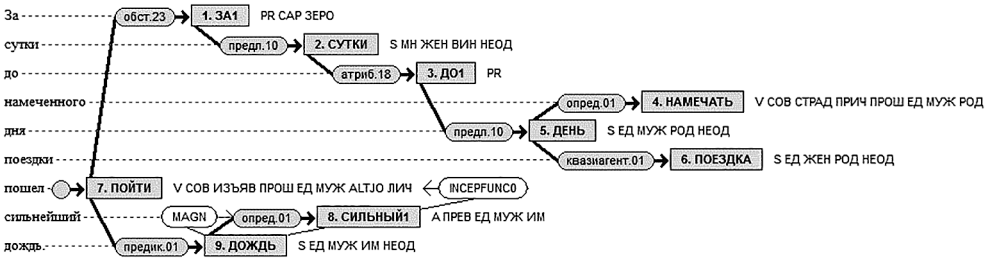

Рис. 4. Синтаксическая структура предложения (4)

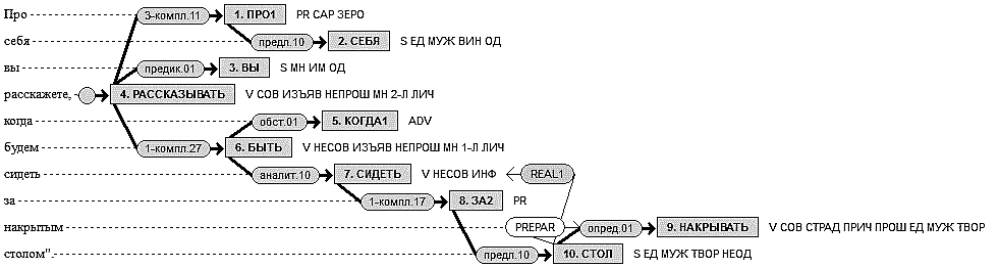

Рис. 5. Синтаксическая структура предложения (5)

Рассмотрим несколько примеров. Выше в предложении (2) узлы 2 (*данным*) и 1 (*по*) связаны ЛФ-отношением AD2-UN (ЛФ, описывающая предлог, сочетание которого с существительным X описывает свойство или состояние второго участника ситуации X). В предложении (3) узлы 22 (*пароходе*) и 21 (*на*) связаны ЛФ-отношением LOC (ЛФ, описывающая предлог, обозначающий нормальную пространственную или временную локализацию чего-либо по отношению к X-у). В предложениях (4) и (5) ниже имеется по два ЛФ-сочетания с одним аргументом. В предложении (4) *сильный дождь* является сочетанием аргумента ДОЖДЬ со значением ЛФ MAGN ‘высокая степень или интенсивность’, *пошел дождь* является сочетанием аргумента со значением ЛФ INCEPFUNC0 ‘начинать иметь место’. В предложении (5) *накрытым столом* представляет собой сочетание аргумента СТОЛ со значением ЛФ PREPAR ‘приготовить X к использованию в соответствии с его назначением’ в форме страдательного причастия, *сидеть за… столом* представляет собой сочетание того же аргумента СТОЛ со значением ЛФ REAL1 ‘использовать X в соответствии с его предназначением’, при этом СИДЕТЬ понимается как значение ЛФ REAL1, а предлог ЗА2 является обязательным контекстом для реализации ЛФ-зависимости. ЛФ-связи, для реализации которых обязателен контекст определенного предлога, отмечаются в структуре записью в овале зеленого цвета, как в предложении (5).

(4) *За сутки до намеченного дня поездки пошел сильнейший дождь.*

(5) *«Про себя вы расскажете, когда будем сидеть за накрытым столом».*

Подробнее о ЛФ-разметке в СинТагРус’е см. в статье [Шеманаева, Фролова 2010].

Поиск по ЛФ доступен онлайн. Однако в настоящее время ЛФ-сочетания в результатах поиска на сайте выделены только в тексте каждого предложения, в структурах же наличие ЛФ-связи и их конкретный вид никак не отражены. Предложение (4), полученное в результате поиска по запросу ЛФ MAGN, будет иметь в онлайновом варианте вид

(4′) За сутки до намеченного дня поездки пошел **сильнейший дождь**.

То же предложение, полученное по запросу ЛФ INCEPFUNC0, будет иметь вид

(4″) За сутки до намеченного дня поездки **пошел** сильнейший **дождь**.

3. Новые типы разметки СинТагРус'а

3.1. Анафорическая разметка

Для целей семантического анализа в лингвистический процессор ЭТАП-4 был включен модуль автоматического разрешения анафоры, находящий для каждого местоимения его антецедент — выражение, к которому отсылает это местоимение и которое, как правило, ему предшествует. Например, в предложении

(6) *Директор*$_i$ *показался в дверях, натягивая макинтош, его*$_i$ *обступили — каждый*$_j$ *со своим*$_j$ *неотложным делом…*

анафорическое местоимение *он* содержит отсылку к антецеденту *директор* (они кореферентны), а местоимение *своим* — к антецеденту *каждый* (анафора без кореферентности, при которой местоимение отсылает к связанной квантором переменной).

Алгоритм поиска антецедентов представляет собой упорядоченную систему правил, написанных на общем для многоцелевого лингвистического процессора ЭТАП-4 языке FORET. Для него предусмотрено два режима работы: автоматический с возможностью последующего редактирования и ручной. Алгоритм запускается при переводе с русского языка на язык семантических представлений и начинает работать после построения синтаксической структуры парсером ЭТАП-4. Анафорические связи являются ориентированными: они направлены от местоимения к существительному (или, в некоторых случаях, прилагательному), которое является синтаксической вершиной антецедента этого местоимения. Кореферентные связи между неместоименными словами на настоящий момент не устанавливаются.

Технически анафорическая (а также микросинтаксическая) разметка СинТагРус'а (см. ниже раздел 3.2) обеспечивается современной версией программной среды «Редактор структур» (StrEd, см. [Iomdin, Sizov 2009]), разработанной в свое время специально для целей создания и поддержки корпуса.

На рисунке 6 показано, как отражаются анафорические связи в Редакторе структур: в поле COREF под полем с морфосинтаксической разметкой представлены пары словоформ (первая — местоимение, вторая — его антецедент).

В основе наших правил разрешения анафоры лежат принципы, перечисленные в книге [Mitkov 2002], а также наши собственные наблюдения (например, о многих

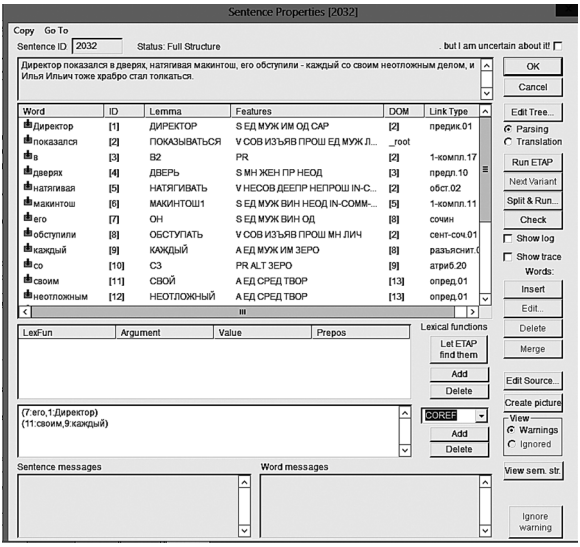

Рис. 6. Представление фразы (6) в Редакторе структур

конструкциях с существительными, неспособными быть антецедентами — см. [Иншакова 2016], о некоторых видах нулевых субъектов). Порядок, в котором применяются правила:

1. Правило разрешения разных видов нулевой анафоры — нулевых субъектов (PRO) причастных, деепричастных и инфинитивных оборотов, относительных, причинных, временных, условных, уступительных конструкций, конструкций меры и степени (*Акция$_i$ оказалась такой успешной, что* PRO$_i$ *была продлена еще на неделю*), предложений с сентенциальными актантами, а также у событийных существительных в некоторых конструкциях (*Его$_i$ обвиняли в* PRO$_i$ *убийстве*). Это правило проводит «фантомные» синтаксические связи между хозяином нулевого местоимения и его контролером.

2. Правила установления антецедентов возвратных местоимений СЕБЯ и СВОЙ, а также возвратного местоимения *друг друга* (связь идет от склоняемого элемента ДРУГ).

3. Правила установления антецедентов относительных местоимений КОТОРЫЙ, КТО, ЧТО2 и ЧЕЙ.

4. Правила установления антецедента «местоимения переключения референции» ТОТ1.

5. Правила установления антецедентов местоимений 3-го лица.

В настоящее время поиск антецедентов местоимений 3-го лица возможен в пределах трех предложений. Он проводится в три этапа:

а) Поиск и маркировка признаком NON-ANTEC таких существительных, которые вообще не могут быть антецедентами местоимений 3-го лица (сами по себе или в определенных конструкциях), например: *Еще один элемент этой конструкции — рама; признание Мадагаскара независимым государством; в присутствии Х-а; сказать в ответ; к сожалению; не в том дело; студия звукозаписи; работать весь день; на 20 лет больше* и мн. др. конструкции (см., например, [Иншакова 2016]).

б) Порождение для каждого местоимения множества гипотетических антецедентов и проведение к ним кореферентных связей. При этом используется морфологическая (согласование), синтаксическая (принципы связывания анафорических местоимений, ограничения на анафору к элементам сочиненной группы), семантическая информация (дескрипторные ограничения на семантику актантов русских

лексем в их моделях управления). Данное правило применяется к одному и тому же местоимению до тех пор, пока не будут отобраны все возможные кандидаты на роль его антецедента.

в) Стирание неверных кореферентных связей. За это отвечают три группы правил:

i. правила, использующие семантическую и онтологическую информацию (онтокорреляты русских лексем из онтологии OntoEtap — см. [Богуславский и др. 2012]);

ii. правила, использующие синтаксическую информацию (невозможные синтаксические конфигурации местоимения и его антецедента, конструкции, в которых происходит переключение референции);

iii. правила, использующие дискурсивную информацию (об относительной дискурсивной выделенности разных кандидатов).

Работа над системой разрешения анафоры описана в статье [Inshakova 2019], там же приведены результаты оценки качества работы системы.

В 2019 году в Лаборатории был создан собственный корпус с анафорической разметкой в формате .tgt. Этот корпус состоит из текстов, входящих в СинТагРус и собранных в 2017–2019 гг. Его объем составляет 6315 предложений (43 текста, бóльшая часть которых — публицистические и художественные, с небольшим количеством новостей и интервью), и 3621 пару «местоимение — антецедент». Он был размечен автоматически, а затем ошибки разметки были исправлены вручную.

На рисунке 7 представлен фрагмент анафорически размеченного корпуса, где кореферентные связи отображаются в столбце Coref. Координата непосредственно перед словоформой означает ее номер во фразе, а дополнительная координата с минусом присутствует у словоформ, которые относятся к предыдущим фразам. Эта координата содержит также смещение относительно текущей фразы (так, номер -1;10 у словоформы *Василий* означает, что она имеет номер 10 в предшествующем предложении; номер -2;2 у словоформы *дочь* указывает на то, что она идет под номером 2 во втором из предшествующих предложений).

Для анафорической разметки корпуса был разработан функционал «Рабочее место» (рис. 8). Он представляет собой диалоговое окно, вызываемое из главного меню Редактора структур. Функционал позволяет проводить кореферентные связи в пределах одного или нескольких предложений как вручную, так и автоматически.

Здесь в окне “Sent. number” задается диапазон предложений, в которых могут строиться кореферентные связи. Если диапазон включает предыдущие предложения, они выводятся в окне “Previous sentences”. В окне, находящемся непосредственно под окном “Previous sentences”, выводится текущее предложение, со словами которого пользователь будет работать. Ниже окна с этим предложением выводится список кореферентных связей, построенных для текущего предложения. В нем указываются координаты в тексте и словоформы для местоимения и антецедента.

Блок “Manual addition” позволяет добавлять разметку вручную. Слово, из которого выходит кореферентная связь, выбирается в выпадающем списке Pronoun.

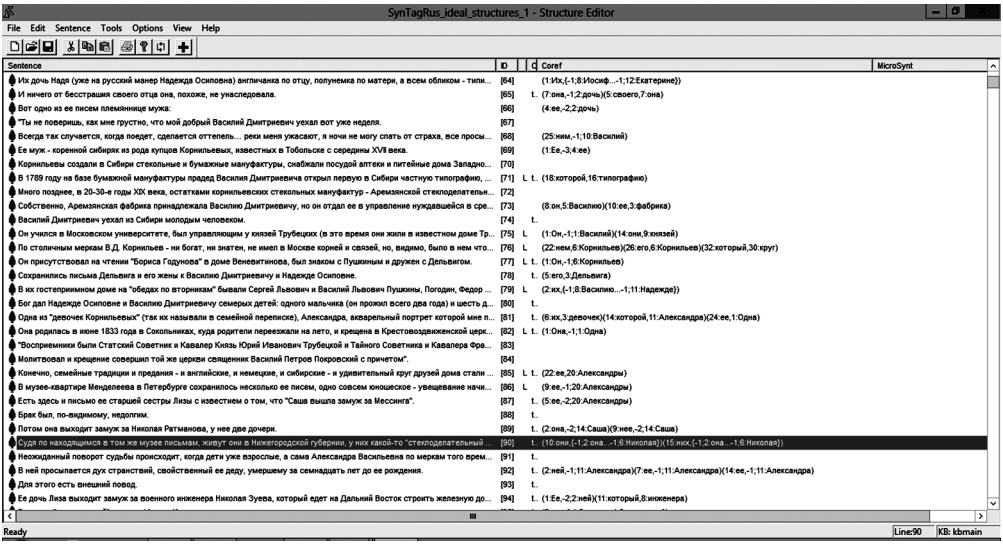

Рис. 7. Фрагмент анафорически аннотированного текста корпуса СинТагРус «А он, мятежный, просит бури…»

Слово-антецедент, в которое связь входит, выбирается в выпадающем списке Begin. После указания слов, входящих в добавляемую кореферентную связь, данная связь добавляется в список связей и запоминается в разметке нажатием кнопки Add. Связи, проведенные вручную, в дальнейшем используются в качестве эталона.

Непосредственная оценка качества работы алгоритма осуществляется с помощью созданной в лаборатории программы, которая автоматически строит кореферентные связи для фраз из корпуса. Построенные связи сравниваются с эталонными, и при обнаружении расхождений выводится сообщение об этом. Также в ходе работы программа запоминает количество правильных (совпадающих с эталоном) связей и общее количество построенных и эталонных связей. На основании этой информации в конце работы вычисляется точность (отношение числа построенных правильных связей к числу всех построенных связей) и полнота (отношение числа построенных правильных связей к числу эталонных связей), используемые для оценки качества системы.

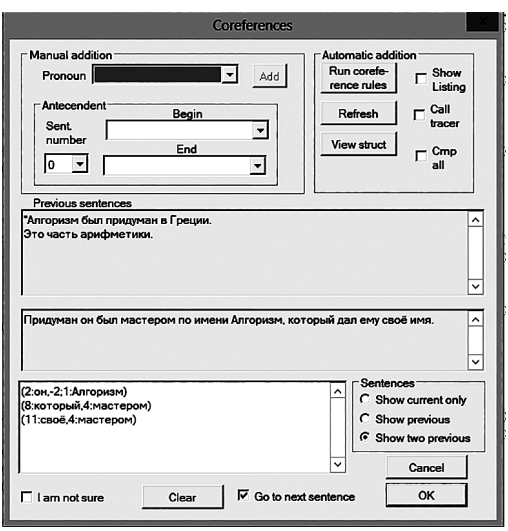

Рис. 8. Рабочее место разметчика

3.2. Микросинтаксическая разметка

Разметка СинТагРус'а микросинтаксическими единицами (то есть, по существу, идентификация таких единиц) представляет собой самое последнее принципиальное обновление структуры корпуса. Она осуществляется недавно, с 2016 г. Этот тип разметки прямо связан с развитием работы по микросинтаксису (из последних публикаций упомянем [Иомдин 2013(а,б), 2014, 2017(а,б), 2018; Iomdin 2015]), в ходе которой стало очевидно, что корпусные данные, включающие информацию о наличии в тексте синтаксически чувствительных фразеологических единиц и их встроенности в структуру предложения, исключительно полезны как для теоретической лингвистики и практической одноязычной и двуязычной лексикографии, так и для решения компьютерно-лингвистических задач, в частности глубокого автоматического семантического анализа. Добавим, что микросинтаксическая разметка корпуса ведется параллельно с созданием Микросинтаксического словаря русского языка, в котором микросинтаксические единицы получают полноценное лексикографическое описание (см. [Iomdin 2016; Иомдин 2017]). Характеристику этого словаря мы в данной статье оставляем в стороне.

Насколько известно авторам, в настоящее время СинТагРус — единственный корпус текстов, в котором эксплицитно размечаются синтаксически чувствительные фразеологические единицы. Разумеется, существуют и активно развиваются одноязычные и параллельные корпусы текстов разных языков, в которых размечаются различного рода коллокации, или multiword expressions (см., например, [Bejček, Straňák 2010; Hnátková et al. 2017; Osenova, Simov 2018], а также обзор релевантных подходов в [Rosén et al. 2016] и [Savary et al. 2017]). Однако в подобной разметке каждая такая единица фиксируется, как правило, без учета ее фразеологичности, почти как единое целое, без представления о том, как именно она интегрируется в синтаксическую структуру предложения, какие в точности индивидуальные значения отдельных слов в нее входят и какие синтаксические связи постулируются между составными элементами единицы. Все это присутствует в микросинтаксической разметке СинТагРус'а.

Микросинтаксические единицы бывают двух типов: нестандартные синтаксические конструкции и синтаксические фраземы.

Нестандартных синтаксических конструкций в русском языке немного: их, по нашим оценкам, несколько десятков. Примерами таких конструкций являются конструкции так называемого малого синтаксиса, такие как инфинитивно-модальные конструкции типа

1) *Y-у X-овать* ('Y должен будет X-овать', ср. *Мне сегодня всю ночь работать*);

2) *Y-у не X-овать* ('Отсутствует перспектива, что Y будет X-овать', ср. *Нашей сборной никогда не пробиться в финал*);

конструкции типа

3) *Y-у не до X-а* ('Y занят более важными делами, чем X, и заявляет, что не будет делать X, считая, что X-м можно пренебречь'; ср. *Коле сейчас не до развлечений*).

Значительное место среди нестандартных синтаксических конструкций занимают разнообразные выражения с обязательным повторением лексических единиц: у каждого такого типа выражений всякий раз своеобразная семантика. Надо сказать, что таких конструкций в русском языке достаточно много; вероятно, их существенно больше, чем, например, в английском, хотя и там их немало[1]. Приведем несколько примеров этих выражений:

4) $X_{инф}$ *не* $X_{личн}$*:* ***Летать не летал, но*** *два раза сидел в салоне Конкорда* [интернет-форум]: (≈ ‘Не летал, но имел место факт «сидел в салоне», что является чем-то более слабым, чем факт «летал»’);

5) $X_{пов}$ *не* $X_{пов}$: *Теперь* ***кричи не кричи, зови не зови*** *— никто не услышит* [В. Солоухин] (≈ ‘Независимо от того, будешь ли ты кричать или нет, будешь ли ты звать или нет, это не приведет к желательному результату...’);

6) $X_{им}$ *есть* $X_{им}$: ***Мальчишки есть мальчишки*** (≈ ‘Мальчишки ведут себя так, как следует ожидать от мальчишек’); ***Жена есть жена*** [А.П. Чехов. Три сестры]; ***закон есть закон*** (≈ ‘закон функционирует так, как следует ожидать от закона’) и т.д.;

7) $X_{им}$ *как* $X_{им}$: ***Машина как машина****, ничего* особенного (≈ ‘Ничем не примечательная машина’); *Нижний Прохора не поразил —* ***город как город*** [В.Я. Шишков];

8) *X так X* (≈ ‘говорящий согласен с предложением X слушающего и готов его выполнить, поскольку говорящему выбор между X и чем-то другим безразличен’): ср. *Хотите еще кофе? — Лучше чаю. —* ***Чай так чай*** [Л. Юзефович]; *Мне вообще казалось неудобным быть выше мамы, и я ей подчеркнуто во всем подчинялся:* ***в лифт — так в лифт****,* ***не заходить*** *в квартиру —* ***так не заходить*** [А. Алексин].

Подавляющая часть микросинтаксических элементов принадлежит к классу синтаксических фразем. Мы понимаем под синтаксическими фраземами такие фразеологические единицы, которые, помимо лексической избирательности и семантической некомпозициональности, обладают заметной синтаксической спецификой (например, управляющими свойствами, отсутствующими у элементов таких единиц: так, единица *руки чешутся* управляет инфинитивом и предлогом *на* (ср. *Чешутся руки написать стихи о состязании двух людей в благородстве* [Ф. Искандер]; *Уж очень у меня на этого Попугайчикова руки чешутся* [А.В. Сухово-Кобылин]), а единица *не по себе* управляет союзом *что* (ср. *И было как-то не по себе, что едешь не в ту сторону, соблазнившись этим неожиданным отпуском от войны* [К. Симонов]).

Некоторые такие единицы тяготеют к цельным словам; ср. единицы типа *всё равно* (по крайней мере в двух из значений — ‘в любом случае’, как в *Скажите, что я всё равно приеду* [Д. Гранин], и ‘равносильно’, как в афоризме М.Л. Гаспарова *«Цзи Юнь говорил: поститься — это все равно что не брать взяток по вторникам и четвергам»*), *всё же, тот же, между тем, между прочим, тем не менее, тем более, только что, пока что, разве что, то и дело, что ли, мало*

[1] Представительный список английских тавтологических конструкций приводится, в частности, в работе [Rhodes 2009].

ли, во что бы то ни стало, в кои веки, союз и частица *как бы* и десятки других. Многие из этих единиц в русском словаре лингвистического процессора ЭТАП-4 представлены как безусловные обороты (см. ниже), их насчитывается порядка двухсот.

В процессе языковых изменений некоторые такие единицы движутся по направлению к единому слову, подобно тому, как это произошло с союзом/наречием *якобы* (которое, согласно данным НКРЯ, могло писаться раздельно вплоть до начала XX в.), или подобно партикулам Т.М. Николаевой [1985, 2008], таким как *ибо, или, либо* и *неужели*, исторически образовавшимся из нескольких отдельных формантов.

Другие синтаксические фраземы свести к цельным словам нельзя. Примером тут является единица *всё равно* в значении ‘безразлично’: в одних случаях форманты *всё* и *равно* стоят рядом; ср. *Мне всё равно, куда ехать отдыхать*, в других они разделены; ср. *Зависеть от царя, зависеть от народа — Не всё ли нам равно?* [А.С. Пушкин]. К этому же подклассу относятся такие разнообразные единицы, как *как быть* (≈ ‘как следует поступать’) в примерах типа *как врачу быть с пациентами, которые отказываются принимать лекарства*, дискурсивное выражение *то ли дело* (ср. *Не понравилась мне Москва: лохматая, кривобокая, — как старушки, горбатые домишки. То ли дело Петербург-щеголь* [А.С. Серафимович]), конструкции типа *черт <дьявол, бес, бог…> знает что <кто, где, когда, зачем…>* и многие другие.

В СинТагРус’е микросинтаксическая разметка распространяется на оба перечисленных вида единиц. Следует отметить, что внесение в корпус такой разметки — весьма нетривиальная задача.

С одной стороны, готовых списков микросинтаксических единиц, которые можно было бы хотя бы с натяжкой считать полными, не существует. Даже традиционные фразеологические словари здесь мало помогают: большинство присутствующих в них идиом не имеют какой-либо синтаксической специфики, а многие микросинтаксические единицы, в том числе близкие к служебным лексическим единицам (составные предлоги, союзы, частицы), во фразеологические словари никогда не попадают.

С другой стороны, бывает далеко не просто отличить микросинтаксическую единицу, которая даже не обязательно принадлежит одной синтаксической группе, от произвольной последовательности слов: ср., например, микросинтаксическую союзную единицу *как бы* в предложении *Боюсь, как бы он не заболел* и случайную цепочку, состоящую из наречия *как* и частицы *бы* в предложении *Как бы ты поступил на моем месте?*

Учитывая недостаток исходных данных, авторы СинТагРус’а используют при его микросинтаксической разметке две дополняющие друг друга стратегии: (1) полный просмотр текста, осуществляемый с целью зафиксировать все возможные микросинтаксические единицы, присутствующие в тексте, и (2) направленный поиск вхождений микросинтаксических единиц, уже известных разметчику. Поисковые запросы при второй стратегии формулируются в терминах линейной

последовательности элементов, характерной для таких единиц, и в терминах фрагментов синтаксических поддеревьев. Эта формулировка осуществляется с помощью специальных правил, написанных на языке FORET. В последнее время, когда микросинтаксическая разметка вводится в новые тексты одновременно с их морфосинтаксической разметкой, разумеется, используется первая стратегия (она применяется аннотатором к каждому предложению).

Идентификация микросинтаксических единиц в корпусе заключается в фиксации имен этих единиц (в случае синтаксических фразем это обычно цепочка слов, возможно, с цифрой, указывающей на конкретное значение фраземы, если она многозначна), а также указании линейных границ, в пределах которых оказываются элементы фиксируемой микросинтаксической единицы. В случае, если та или иная единица не представляет собой непрерывной последовательности идущих друг за другом слов, а располагается менее плотно, информация о лексическом составе единицы может оказаться недостаточно определенной (ср., например, представление синтаксической фраземы *как быть* в конструкциях типа *Не знаю уж, как и быть мне с тобой*… [М.Е. Салтыков-Щедрин] и даже, казалось бы, совсем уж неделимой единицы *несмотря на* в контекстах типа *В крещенский праздник, несмотря ни на какие морозы, купаются сотни горожан* [О. Белякова]: здесь линейные отрезки содержат посторонние микросинтаксическим единицам элементы *и* и *ни*).

В настоящее время микросинтаксическая разметка представляет собой особое поле в XML-файле, соответствующем размечаемому тексту. В этом поле фиксируется имя микросинтаксической единицы и линейный отрезок, в пределах которого она располагается. Графическое представление этого поля, порождаемое Редактором структур, дается на рисунке 9.

Микросинтаксическая разметка корпуса производится в основном вручную. Исключение составляют ситуации, когда в словаре системы используются в качестве лексических единиц неоднословные токены (так называемые «безусловные обороты»), которые в подавляющем большинстве случаев представляют собой микросинтаксические единицы (конкретнее, синтаксические фраземы). Примерами таких единиц могут служить обороты *а именно*, *а также*, *а то и*, *без устали*, *бок о бок*, *в виду* (для представления выражений *иметь в виду* и *иметься в виду*), *в качестве*, *в конечном счете*, *в конце концов*, *в настоящее время*, *в противном случае*, *в связи с*, *в силу*, *в складчину*, *в соответствии с*, *в состоянии*, *в течение*, *в то время как*, *в частности*, *во время*, *во избежание*, *во что бы то ни стало*, *во языцех*, *вплоть до*, *вряд ли* и др. Их можно автоматически квалифицировать как таковые, конечно, с последующей экспертной редакцией. Скажем, единица *как только* практически всегда представляет собой единый союз, однако изредка встречаются и ситуации, когда эта цепочка слов должна трактоваться по отдельности; ср. *Как только тебе не стыдно?* Точно так же последовательность *всё же* в подавляющем числе случаев выступает как единая микросинтаксическая единица — частица; ср. *И всё же мы часто ссорились* [С. Довлатов], но изредка представляет два разных слова; ср. *Не всё же разглагольствовать о том, каким должен быть хороший*

человек, пора и стать им [С. Смирнов]. В подобных ситуациях микросинтаксическая интерпретация таких последовательностей оказывается ошибочной и удаляется из разметки.

Иногда работа с такими единицами при разметке корпуса приводит к неожиданным и даже курьезным результатам. Так, у той же синтаксической фраземы *всё же* имеется однословный, почти полностью синонимичный аналог (почти полный синоним) *всё-таки*, а у этого последнего имеется трехсловный вариант *всё ж таки* (который является вполне частотным: в основном корпусе НКРЯ имеется 670 его вхождений), и его разумно признать безусловным оборотом, в отличие от спорной, как мы только что видели, единицы *всё же*.

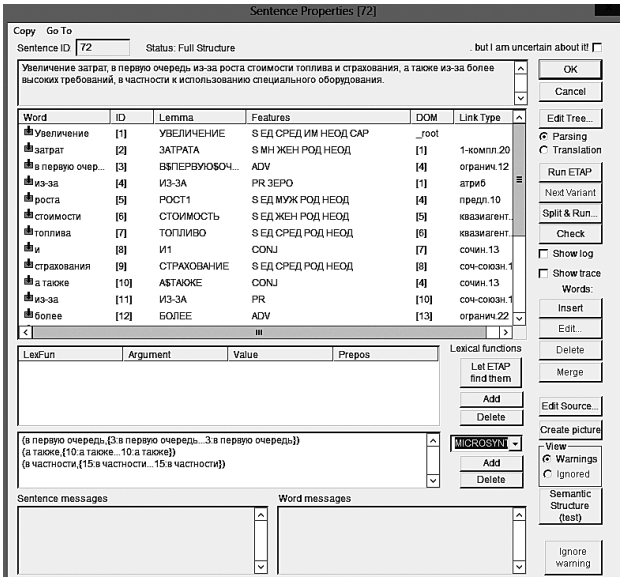

Рис. 9. Представление фразы, содержащей три микросинтаксические единицы — *в первую очередь, а также* и *в частности*

Легко убедиться, что микросинтаксически размеченный корпус текстов представляет собой ценнейший лингвистический материал для исследования русской, а возможно, и сопоставительной фразеологии.

В частности, по соотношению количества истинных и ложно-положительных вхождений той или иной нестандартной синтаксической конструкции или синтаксической фраземы можно судить о ее фразеологической силе: если в подкорпусе предложений, содержащих такие вхождения, их статус достаточно однороден и истинные вхождения микросинтаксической единицы существенно преобладают над ложными, можно считать, что данная единица обладает высокой степенью фразеологичности. Практическим следствием такой оценки может быть утверждение данной единицы безусловным оборотом, даже если теоретически можно допустить, что в реальном тексте могут присутствовать последовательности элементов некоторой единицы, которые ее, однако, не формируют.

Показательным примером является синтаксическая фразема *в первую очередь*. Нетрудно заметить, что в русском тексте последовательность этих словоформ может фигурировать как свободное сочетание (как во фразе *Не становись в первую очередь, она очень медленно движется*) или даже как случайное соположение слов, не образующих смыслового единства (ср. *Работали две кассы, в первую очередь стояла огромная, а во вторую почему-то небольшая*). Тем не менее анализ микросинтаксической разметки СинТагРус’а показывает, что все 125 вхождений

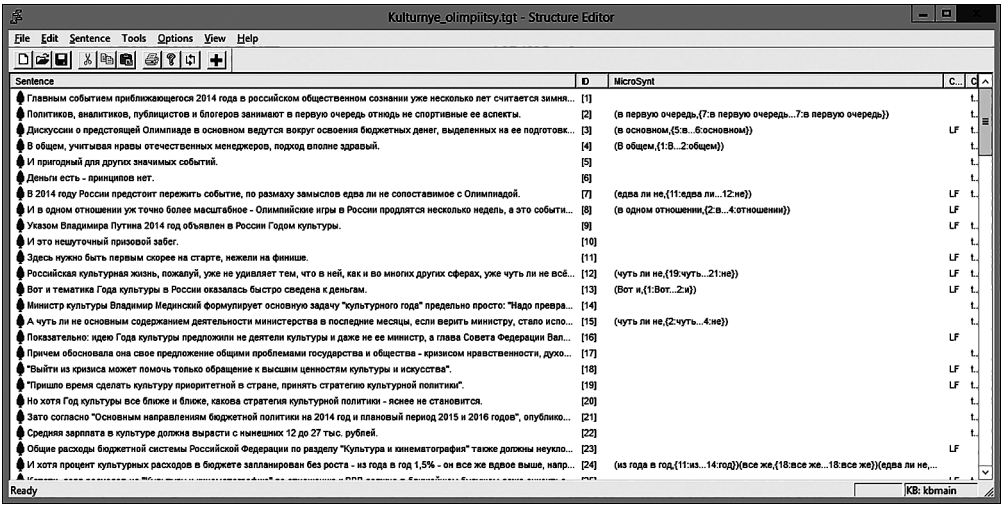

Рис. 10. Фрагмент микросинтаксически аннотированного текста корпуса СинТагРус «Культурные олимпийцы». Этот текст содержит 133 предложения, из которых в 34 предложениях (25%) присутствует хотя бы один микросинтаксический элемент

последовательности словоформ *в*, *первую* и *очередь* представляют собой именно синтаксическую фразему.

Как показывают предварительные статистические оценки, в среднем микросинтаксические единицы присутствуют в четверти предложений типичного текста СинТагРус'а. На рисунке 10 приводится фрагмент одного такого текста.

В сегодняшней версии СинТагРус'а подкорпус, содержащий микросинтаксическую разметку, составляет более 2500 предложений. Общее число встречающихся в нем разных микросинтаксических единиц — около 300.

Более детальное описание микросинтаксической разметки СинТагРус'а можно найти в работах [Маракасова, Иомдин 2016] и [Iomdin 2017].

4. Использование корпуса СинТагРус

Корпус СинТагРус, как и другие глубоко аннотированные корпусы текстов, активно используется для целей теоретико-лингвистических исследований (в первую очередь, разумеется, в области синтаксиса) и для практической лексикографии.

Кроме того, корпус применяется в компьютерно-лингвистических задачах в качестве источника лингвистических данных. Конкретные задачи могут быть самыми разными — от создания автоматических парсеров на основе машинного обучения (см., например, опыт создания Maltparser'а для русского языка в [Nivre et al. 2008]) до использования корпусной статистики при оптимизации правиловых парсеров и для регрессионного тестирования синтаксической модели, лежащей в основе СинТагРус'а.

Добавим, что как среди компьютерных лингвистов, так и среди лингвистов-теоретиков большим спросом пользуется оффлайновая версия СинТагРус'а:

к настоящему времени образовательным и научным учреждениям и отдельным исследователям выдано свыше 100 лицензий на некоммерческое использование этого ресурса.

Особого упоминания заслуживает сравнительно новая практика использования синтаксически размеченных корпусов текстов для создания унифицированных производных корпусов текстов.

Так, в рамках многоязычного некоммерческого проекта Universal Dependencies (http://universaldependencies.org/) СинТагРус был преобразован в формат универсальных зависимостей — гармонизированную систему представления синтаксической структуры в виде деревьев зависимостей, которая в достаточной степени подходит для описания разноструктурных языков. На основании этого формата был создан единый стандарт, с помощью которого аннотируются корпусы текстов на разных языках, если они предусматривают синтаксическую разметку (подробнее о нем см. [Nivre 2015] и [Lyashevskaya et al. 2016]). До настоящего времени не разрабатывались новые корпусы, которые с самого начала размечались бы по данному стандарту, но на 2018 г. (релиз от 1 июля) в этот стандарт преобразовано 122 корпуса на 71 языке [Nivre et al. 2018]. В число этих корпусов входит четыре корпуса русского языка — три небольших экспериментальных корпуса, несопоставимых с СинТагРус'ом по объему, и корпус, полученный из СинТагРус'а (без непосредственного участия его разработчиков). Этот корпус фигурирует на портале Universal Dependencies под именем UD_Russian-SynTagRus и соответствует состоянию СинТагРус'а на середину 2016 г.

Литература

Апресян Ю.Д., Богуславский И.М., Иомдин Б.Л., Иомдин Л.Л., Санников А.В., Санников В.З., Сизов В.Г., Цинман Л.Л. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003–2005. М. : Индрик, 2005. С. 193–214.

Богуславский И.М., Диконов В.Г., Тимошенко С.П. Онтология для поддержки задач извлечения смысла из текста на естественном языке // Информационные технологии и системы 2012 (ИТиС'2012). Труды 35-й междисциплинарной школы-конференции ИППИ РАН. Петрозаводск, 2012. С. 152–161.

Богуславский И.М., Иомдин Л.Л., Валеев Д.Р., Сизов В.Г. Синтаксический анализатор системы ЭТАП и его оценка с помощью глубоко размеченного корпуса русских текстов // Международная конференция «Корпусная лингвистика — 2008». СПб., 2008а. С. 56–74.

Богуславский И.М., Иомдин Л.Л., Митюшин Л.Г., Сизов В.Г. Длина синтаксических связей в русском аннотированном корпусе // Международная конференция «Корпусная лингвистика — 2008». СПб., 2008б. С. 75–82.

Дяченко П.В., Иомдин Л.Л., Лазурский А.В., Митюшин Л.Г., Подлесская О.Ю., Сизов В.Г., Фролова Т.И., Цинман Л.Л. Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) // Национальный корпус

русского языка: 10 лет проекту. Труды Института русского языка им. В.В. Виноградова. Вып. 6. М., 2015. С. 272–299.

Жолковский А.К., Мельчук И.А. О семантическом синтезе // Проблемы кибернетики. Вып. 19. М. : Наука, 1967. С. 177–238.

Иншакова Е.С. Разрешение синтаксической местоименной анафоры в системе ЭТАП-3 // Информационные технологии и системы 2016 (ИТиС'2016). Труды 40-й междисциплинарной школы-конференции ИППИ РАН. СПб., 2016. С. 420–429.

Иомдин Л.Л. Читать не читал, но…: об одной русской конструкции с повторяющимися словесными элементами // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». Вып. 12 (19). Т. 1. М. : Изд-во РГГУ, 2013а. С. 297–310.

Иомдин Л.Л. Некоторые микросинтаксические конструкции в русском языке с участием слова *что* в качестве составного элемента // Јужнословенски филолог. Т. LXIX. Белград: Институт сербского языка Сербской академии наук и искусств, 2013б. С. 137–147.

Иомдин Л.Л. Хорошо меня там не было: синтаксис и семантика одного класса русских разговорных конструкций // Grammaticalization and Lexicalization in the Slavic Languages. По материалам Международного симпозиума «Грамматикализация и лексикализация в славянских языках», 11–14 ноября 2011 г. München ; Berlin ; Washington D.C. : Verlag Otto Sagner, 2014. Bd. 55. S. 423–436.

Иомдин Л.Л. Как нам быть с конструкциями типа *как быть?* // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». Вып. 16 (23). Т. 1. М. : Изд-во РГГУ, 2017а. С. 161–176.

Иомдин Л.Л. Между синтаксической фраземой и синтаксической конструкцией. Нетривиальные случаи микросинтаксической неоднозначности // SLAVIA, časopis pro slovanskou filologii, ročník 86, sešit 2–3. 2017б. S. 230–243.

Иомдин Л.Л. Еще раз о микроконструкциях, сформированных служебными словами: *то и дело* // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». Вып. 17 (24). М. : Изд-во РГГУ, 2018. С. 267–283.

Маракасова А.А., Иомдин Л.Л. Микросинтаксическая разметка в корпусе русских текстов СинТагРус // Информационные технологии и системы 2016 (ИТиС'2016). Труды 40-й междисциплинарной школы-конференции ИППИ РАН. СПб., 2016. С. 445–449.

Мельчук И.А. Опыт теории лингвистических моделей «Смысл ↔ Текст». М. : Наука, 1974. 314 с. (2-е изд.: 1999. 346 с.)

Николаева Т.М. Функции частиц в высказывании (на матер-иале славянских языков). М. : Наука, 1985. 170 с.

Николаева Т.М. Непарадигматическая лингвистика (История «блуждающих частиц»). М. : Языки славянской культуры, 2008. 689 с.

Шеманаева О.Ю., Фролова Т.И. Лексико-функциональная разметка текстов в СинТагРус // Информационные технологии и системы 2010 (ИТиС'10). Труды

33-й конференции молодых ученых и специалистов ИППИ РАН. М. : ИППИ, 2010. С. 320–324.

Apresjan Ju., Boguslavsky I., Iomdin L., Lazursky A., Sannikov V., Sizov V., Tsinman L. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT // MTT 2003, First International Conference on Meaning-Text Theory. P., 2003. P. 279–288.

Bejček E., Straňák P. Annotation of Multiword Expressions in the Prague Dependency Treebank // Language Resources and Evaluation. 2010. Vol. 44. No. 1–2. P. 7–21.

Boguslavsky I. SynTagRus — a Deeply Annotated Corpus of Russian // Les èmotions dans le discours. Emotions in Discourse / eds. Peter Blumenthal, Iva Novakova, Dirk Siepmann. Peter Lang Edition, 2014. P. 367–381.

Boguslavsky I., Grigorieva S., Grigoriev N., Kreidlin L., Frid N. Dependency Treebank for Russian: Concept, Tools, Types of Information // Proc. of the 18th International Conference on Computational Linguistics (COLING 2000). San Francisco : Kaufmann, 2000. P. 987–991.

Hnátková M., Petkevič V., Skoumalová H. Multiword Expressions in Czech: Between Lexicon and Grammar // *Corpus Linguistics* — 2017. St. Petersburg, St. Petersburg State University: 2017. P. 36–42.

Inshakova E. S. An anaphora resolution system for Russian based on ETAP-4 linguistic processor. Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoi Mezhdunarodnoi konferentsii 'Dialog' [Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"]. Iss. 18 (25). Moscow: RSUH Publ., 2019. P. 239–251.

Iomdin L. The Challenge of Treating Collocations // International Journal of Lexicography. 2015. Vol. 28. No. 3. P. 376–384.

Iomdin L. Microsyntactic Phenomena as a Computational Linguistics Issue // Grammar and Lexicon: Interactions and Interfaces. Proc. of the Workshop. Osaka, 2016. P. 8–18. Available at: http://aclweb.org/anthology/W/W16/W16-38.pdf.

Iomdin L. Microsyntactic Annotation of Corpora and its Use in Computational Linguistics Tasks // Jazykovedný časopis, ročník 86, 2017, čislo 2. S. 169–178.

Iomdin L., Sizov V. Structure Editor: a Powerful Environment for Tagged Corpora // MONDILEX Fifth Open Workshop. Ljubljana, 2009. P. 1–12.

Lyashevskaya O., Droganova K., Zeman D., Alexeeva M., Gavrilova T., Mustafina N., Shakurova E. Universal Dependencies for Russian: A New Syntactic Dependencies Tagset // NRU HSE. Series WP BRP «Linguistics». No. 44. 2016.

Mitkov R. Anaphora resolution. Longman, 2002. 220 p.

Nivre J. Towards a Universal Grammar for Natural Language Processing // Computational Linguistics and Intelligent Text Processing (CICLing 2015). Part 1. Springer, 2015. P. 3–16.

Nivre J., Abrams M., Agić Ž. et al. Universal Dependencies 2.2 // LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL). Faculty of Mathematics and Physics, Charles University, 2018. Available at: http://hdl.handle.net/11234/1-2837.

Nivre J., Boguslavsky I., Iomdin L. Parsing the SynTagRus Treebank of Russian // Proc. of the 22nd International Conference on Computational Linguistics (COLING'08). Manchester, August 18–22 2008 / eds. D. Scott, H. Uszkoreit. Vol. 1. 2008. P. 641–648.

Osenova P., Simov K. Modelling multiword expressions in a parallel Bulgarian-English newsmedia corpus // Multiword expressions: Insights from a multi-lingual perspective. Phraseology and Multiword Expressions / eds. M. Sailer, S. Markantonatou. Berlin : Language Science Press, 2018. P. 247–270.

Rhodes R. Tautological constructions in English … and beyond // Presented to the Syntax and Semantics Circle, UCB, 2009.

Available at: http://linguistics.berkeley.edu/~russellrhodes/pdfs/syntax_circle_taut_qp.pdf.

Rosén V., De Smedt K., Losnegaard G.S., Bejček E., Savary A., Osenova P. MWEs in Treebanks: From Survey to Guidelines // Proc. of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. 2016. P. 2323–2330.

Savary A., Sangati F., Candito M. et al. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions // Proc. of the 13th Workshop on Multiword Expressions (MWE 2017), Valencia, Spain, 4 April 2017. P. 31–47.

[1]Ye.S. Inshakova, [2]L.L. Iomdin, [3]L.G. Mityushin,
[4]V.G. Sizov, [5]T.I. Frolova, [6]L.L. Tsinman
[123456]A.A. Kharkevich Institute for Information Transmission Problems,
Russian Academy of Sciences,
[2]Russian State University for the Humanities
(Russia, Moscow)
[1]e.s.inshakova@gmail.com, [2]iomdin@gmail.com, [3]mit@iitp.ru,
[4]sizov@iitp.ru, [5]tfrolova@gmail.com, [6]llcinman@gmail.com

THE SYNTAGRUS TODAY

The paper describes the current state of the SynTagRus corpus composed of Russian texts tagged with morphosyntactic structures. At certain points of the corpus development, additional kinds of tagging were introduced, namely lexical-semantic, lexical-functional, anaphoric and microsyntactic tagging.

The morphosyntactic tagging of a sentence includes morphological structures of all the words and the syntactic structure of the sentence in the form of a dependency tree, in accordance with I. Mel'čuk and A. Zholkovsky's "Meaning — Text" model. Lexical-semantic tagging implies that each word is assigned a corresponding entry of the Russian combinatorial dictionary. Lexical-functional tagging means finding the phrases in the text which can be interpreted in terms of lexical functions. Anaphoric tagging results in marking the antecedents of pronouns. Microsyntactic tagging identifies syntactic phrasemes and certain nonstandard syntactic constructions occuring in the text.

Tagging of a new text is performed in several stages. First, the text is processed by the multifunctional linguistic processor ETAP-4 which makes morphosyntactic and lexical-semantic tagging in automatic mode. Then the output of the processor is checked and possibly corrected by specially trained annotators. After that, ETAP-4 uses the morpho-syntactic structures of the text to make lexical-functional and anaphoric tagging. Finally, the annotators check these types of tagging and manually perform microsyntactic tagging.

The SynTagRus corpus may be used in theoretical linguistic research as well as practical lexicography. The corpus statistics may help optimize decision making in various automatic text processing procedures. Another very promising possibility is to use the SynTagRus data as input for the modern machine learning systems.

Key words: SynTagRus, syntactically tagged corpus, corpus of Russian texts, dependency grammar, lexical functions, pronoun antecedents, microsyntax, ellipsis

References

Apresjan Ju., Boguslavsky I., Iomdin L., Lazursky A., Sannikov V., Sizov V., Tsinman L. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. *MTT 2003, First International Conference on Meaning-Text Theory*. Paris, 2003, pp. 279–288.

Apresjan Ju.D., Boguslavsky I.M., Iomdin B.L., Iomdin L.L., Sannikov A.V., Sannikov V.Z., Sizov V.G., Cinman L.L. [Syntactically and semantically tagged corpus of Russian: state of the art and prospects]. *Natsional'nyi korpus russkogo yazyka: 2003–2005* [The Russian National Corpus: 2003–005. Results and Prospects]. Moscow, Indrik Publ., 2005, pp. 193–214. (In Russ.)

Bejček E., Straňák P. Annotation of Multiword Expressions in the Prague Dependency Treebank. *Language Resources and Evaluation*, 2010, vol. 44, no. 1–2, pp. 7–21.

Boguslavsky I. SynTagRus — a Deeply Annotated Corpus of Russian. *Les èmotions dans le discours. Emotions in Discourse.* Eds. Peter Blumenthal, Iva Novakova, Dirk Siepmann. Peter Lang Edition, 2014, pp. 367–381.

Boguslavsky I.M., Dikonov V.G., Timoshenko S.P. [Ontology for supporting the tasks of meaning extraction from texts in natural languages]. *Informatsionnye tekhnologii i sistemy 2012 (ITiS'2012). Trudy 35-i mezhdistsiplinarnoi shkoly-konferentsii IPPI RAN* [Information Technologies and Systems 2012 (ITiS'2012). Proc. of the 35th Interdisciplinary School-Conference of IITP RAS]. Petrozavodsk, 2012, pp. 152–161. (In Russ.)

Boguslavsky I., Grigorieva S., Grigoriev N., Kreidlin L., Frid N. Dependency Treebank for Russian: Concept, Tools, Types of Information. *Proc. of the 18th International Conference on Computational Linguistics (COLING 2000)*. San Francisco, Kaufmann, 2000, pp. 987–991.

Boguslavsky I.M., Iomdin L.L., Mitjushin L.G., Sizov V.G. [The length of syntactic links in the Russian tagged corpus]. *Mezhdunarodnaya konferentsiya "Korpusnaya lingvistika — 2008"* [Proc. of the International Conference "Corpus Linguistics — 2008"]. St. Petersburg, 2008b, pp. 75–82. (In Russ.)

Boguslavsky I.M., Iomdin L.L., Valeev D.R., Sizov V.G. [A syntactic analyzer of the ETAP system and its evaluation with the help of a deeply annotated corpus of Russian texts]. *Mezhdunarodnaya konferentsiya "Korpusnaya lingvistika — 2008"* [Proc. of the International Conference "Corpus Linguistics — 2008"]. St. Petersburg, 2008a, pp. 56–74. (In Russ.)

Dyachenko P.V., Iomdin L.L., Lazursky A.V., Mityushin L.G., Podlesskaya O.Yu., Sizov V.G., Frolova T.I., Tsinman L.L. [A deeply annotated corpus of Russian texts (SynTagRus): contemporary state of affairs]. *Natsional'nyi korpus russkogo yazyka: 10 let proektu. Trudy Instituta russkogo yazyka im. V.V. Vinogradova. Vyp. 6* [The Russian National Corpus: 10 Years of the Project. Proc. of the V.V. Vinogradov Russian Language Institute. Iss. 6]. Moscow, 2015, pp. 272–299. (In Russ.)

Hnátková M., Petkevič V., Skoumalová H. Multiword Expressions in Czech: Between Lexicon and Grammar. *Proc. of the Conference "Corpus Linguistics — 2017".* St. Petersburg, St. Petersburg State University, 2017, pp. 36–42.

Inshakova E.S. [Resolution of syntactic pronominal anaphora in the ETAP-3 system]. *Informatsionnye tekhnologii i sistemy 2016 (ITiS'2016). Trudy 40-i mezhdistsiplinarnoi shkoly-konferentsii IPPI RAN* [Information Technologies and Systems 2016 (ITiS'2016). Proc. of the 40th Interdisciplinary School-Conference of IITP RAS]. St. Petersburg, 2016, pp. 420–429. (In Russ.)

Inshakova E.S. An anaphora resolution system for Russian based on ETAP-4 linguistic processor. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoi Mezhdunarodnoi konferentsii 'Dialog'* [Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"]. Iss. 18 (25). Moscow, RSUH Publ., 2019, pp. 239–251.

Iomdin L.L. ["Chitať ne chital, no...": on one Russian construction with repeating items]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoi Mezhdunarodnoi konferentsii "Dialog"* [Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"]. Iss. 12 (19), vol. 1. Moscow, RSUH Publ., 2013a, pp. 297–310. (In Russ.)

Iomdin L.L. [Certain microsyntactic constructions in Russian which contain the word "chto" as a constituent element]. Južnoslovenski filolog, vol. LXIX. Beograd, 2013b, pp. 137–147. (In Russ.)

Iomdin L.L. [Good thing I wasn't there: syntax and semantics of a class of Russian colloquial constructions]. *Grammaticalization and lexicalization in the Slavic languages. Papers from the 36th meeting of the commission on the grammatical structure of the Slavic languages of the International committee of Slavists.* München/Berlin/Washington D.C., Verlag Otto Sagner, 2014, band 55, pp. 423–436. (In Russ.)

Iomdin L. The Challenge of Treating Collocations. *International Journal of Lexicography*, 2015, vol. 28, no. 3, pp. 376–384.

Iomdin L. Microsyntactic Phenomena as a Computational Linguistics Issue. *Grammar and Lexicon: Interactions and Interfaces. Proc. of the Workshop.* Osaka, 2016, pp. 8–18. Available at: http://aclweb.org/anthology/W/W16/W16-38.pdf.

Iomdin L. L. [What to do about constructions like “what to do”?]. *Komp’yuternaya lingvistika i intellektual’nye tekhnologii. Po materialam ezhegodnoĭ Mezhdunarodnoĭ konferentsii “Dialog”* [Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”]. Iss. 16 (23), vol. 1. Moscow, RSUH Publ., 2017a, pp. 161–176. (In Russ.)

Iomdin L. L. [Between a syntactic phraseme and a syntactic construction. Nontrivial cases of microsyntactic ambiguity]. SLAVIA, časopis pro slovanskou filologii, ročník 86, sešit 2–3. 2017b, pp. 230–243. (In Russ.)

Iomdin L. Microsyntactic Annotation of Corpora and its Use in Computational Linguistics Tasks. *Jazykovedný časopis, ročník 86, čislo 2*, 2017c, pp. 169–178.

Iomdin L. L. [Once more about microconstructions formed with function words: “to i delo”]. *Komp’yuternaya lingvistika i intellektual’nye tekhnologii. Po materialam ezhegodnoĭ Mezhdunarodnoĭ konferentsii “Dialog”* [Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”]. Iss. 17 (24). Moscow, RSUH Publ., 2018, pp. 267–283. (In Russ.)

Iomdin L., Sizov V. Structure Editor: a Powerful Environment for Tagged Corpora. *MONDILEX Fifth Open Workshop*. Ljubljana, 2009, pp. 1–12.

Lyashevskaya O., Droganova K., Zeman D., Alexeeva M., Gavrilova T., Mustafina N., Shakurova E. Universal Dependencies for Russian: A New Syntactic Dependencies Tagset. *NRU HSE. Series WP BRP “Linguistics”*, 2016, no. 44.

Marakasova A. A., Iomdin L. L. [Microsyntactic tagging in the SynTagRus corpus of Russian texts]. *Informatsionnye tekhnologii i sistemy 2016 (ITiS’2016). Trudy 40-i mezhdistsiplinarnoi shkoly-konferentsii IPPI RAN* [Information Technologies and Systems 2016 (ITiS’2016). Proc. of the 40th Interdisciplinary School-Conference of IITP RAS]. St. Petersburg, 2016, pp. 445–449. (In Russ.)

Mel’čuk I. A. Opyt teorii lingvisticheskikh modelei “Smysl ↔ Tekst” [The theory of linguistic models of the Meaning — Text type]. Moscow, Nauka Publ., 1974. 314 p. (In Russ.)

(2nd ed.: 1999. 346 p.)

Mitkov R. Anaphora resolution. Longman, 2002. 220 p.

Nikolaeva T. M. *Funktsii chastits v vyskazyvanii (na materiale slavyanskikh yazykov)* [The function of particles in an utterance (based on the materials of *Slavonic languages*)]. Moscow, Nauka Publ., 1985. 170 p. (In Russ.)

Nikolaeva T. M. *Neparadigmaticheskaya lingvistika (Istoriya “bluzhdayushchikh chastits”)* [*Non-paradigmatic* linguistics (History of the “wandering particles”)]. Moscow, LRC Publ., 2008. 689 p. (In Russ.)

Nivre J. Towards a Universal Grammar for Natural Language Processing. *Proc. of Computational Linguistics and Intelligent Text Processing (CICLing 2015).* Part 1. Springer, 2015, pp. 3–16.

Nivre J., Abrams M., Agić Ž. et al. Universal Dependencies 2.2. *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL)*. Faculty of Mathematics and Physics, Charles University, 2018. Available at: http://hdl.handle.net/11234/1-2837.

Nivre J., Boguslavsky I., Iomdin L. Parsing the SynTagRus Treebank of Russian. *Proc. of the 22nd International Conference on Computational Linguistics (COLING'08)*. Manchester, August 18–22 2008. Eds. D. Scott, H. Uszkoreit, 2008, vol. 1, pp. 641–648.

Osenova P., Simov K. Modelling multiword expressions in a parallel Bulgarian-English newsmedia corpus. *Multiword expressions: Insights from a multi-lingual perspective. Phraseology and Multiword Expressions*. Eds. M. Sailer, S. Markantonatou. Berlin, Language Science Press, 2018, pp. 247–270.

Rhodes R. Tautological constructions in English … and beyond. *Presented to the Syntax and Semantics Circle*, *UCB, 2009*.

Available at: http://linguistics.berkeley.edu/~russellrhodes/pdfs/syntax_circle_taut_qp.pdf.

Rosén V., De Smedt K., Losnegaard G. S., Bejček E., Savary A., Osenova P. MWEs in Treebanks: From Survey to Guidelines. *Proc. of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia, 2016, pp. 2323–2330.

Savary A., Sangati F., Candito M. et al. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. *Proc. of the 13th Workshop on Multiword Expressions (MWE 2017), Valencia, Spain, 4 April 2017*, pp. 31–47

Shemanaeva O.Yu., Frolova T. I. [Tagging with lexical functions in SynTagRus]. *Informacionnye tehnologii i sistemy 2010 (ITiS'10). Trudy 33-i Konferencii molodykh uchenykh i spetsialistov IPPI RAN* [Information Technologies and Systems 2010. Proc. of the 33rd Conference of Young Scientists and Specialists of IITP RAS]. Moscow, IITP, 2010, pp. 320–324. (In Russ.)

Zholkovsky A. K., Mel'čuk I. A. [On semantic synthesis]. *Problemy kibernetiki. Vyp. 19* [Problems of cybernetics. Iss. 19]. Moscow, Nauka Publ., 1967, pp. 177–238. (In Russ.)