DOI: 10.31912/pvrli-2024.4.9

И. М. Богуславский, П. В. Дяченко, Е. С. Иншакова, Л. Л. Иомдин, А. В. Лазурский, Л. Г. Митюшин, А. А. Мовсесян, И. П. Рыгаев, В. Г. Сизов, С. П. Тимошенко, Т. И. Фролова, А. В. Чага

ИППИ РАН им. А. А. Харкевича
(Москва)

igor.m.boguslavsky@gmail.com, pavel.v.djachenko@gmail.com, e.s.inshakova@gmail.com, iomdin@gmail.com, lazursky@mail.ru, lmityushin@gmail.com, derise@iitp.ru, irygaev@jent.ru, victor.sizov@gmail.com, nyrestein@gmail.com, tfrolova@gmail.com, chagachaga@gmail.com

СОВРЕМЕННОЕ СОСТОЯНИЕ КОРПУСА СИНТАГРУС[1]

Предлагается описание основных особенностей и опций многосторонне размеченного корпуса русских текстов СинТагРус. Корпус был разработан в ИППИ РАН им. А. А. Харкевича и в настоящее время выступает как один из подкорпусов НКРЯ, где он называется «Синтаксическим корпусом». Излагаются основные подходы к выбору текстов для корпуса и к их метаразметке, лингвистические принципы, лежащие в основе разметки разных типов — морфологической, синтаксической, лексико-семантической, лексико-функциональной, эллиптической, микросинтаксической, кореферентной и темпоральной. Приводятся статистические данные, характеризующие различные аспекты СинТагРуса и его фрагментов. СинТагРус является корпусом со стопроцентно дизамбигуированной на всех уровнях разметкой: в статье описываются очевидные достоинства такой разметки и в то же время отмечаются трудности, связанные с необходимостью всегда принимать определенные решения и выбирать единственные варианты разметки даже в тех случаях, когда языковой материал принципиально допускает неединственное лингвистическое описание. Значительное внимание уделяется описанию некоторых различий, существующих между СинТагРусом и основными подкорпусами НКРЯ — разделению материала по частям речи, различным морфологическим решениям, принятыми в СинТагРусе и НКРЯ (таким, как трактовка отдельных морфологических категорий — вида и залога глагола, некоторых падежей существительных и др.).

[1] Данная работа выполнена в рамках гранта Министерства науки и высшего образования РФ № 075-15-2020-793. Авторы выражают благодарность Министерству за поддержку.

Ключевые слова: СинТагРус, синтаксический корпус, морфосинтаксическая разметка, лексическая разметка, эллиптическая разметка, микросинтаксис, референтная разметка, темпоральная разметка

1. Общая информация о корпусе

В 2023 году синтаксически размеченный корпус русских текстов, известный как СинТáгРус (или SynTagRus, Syntactically Tagged Russian text corpus), отмечает 25-летие своего существования. Работа над корпусом началась в 1998 году в Лаборатории компьютерной лингвистики ИППИ РАН им. А. А. Харкевича и поддерживалась грантами РФФИ, РНФ и РГНФ. Ход работы над корпусом и различные теоретические и практические аспекты его функционирования и использования отражены в ряде публикаций его авторов (см., в частности, [Апресян и др. 2005], [Boguslavsky et al. 2000], [Богуславский и др. 2008], [Iomdin, Sizov 2009], [Шеманаева, Фролова 2010], [Boguslavsky 2014], [Дяченко и др. 2015], [Iomdin 2016], [Маракасова, Иомдин 2016], [Iomdin 2017], [Тимошенко и др. 2021], [Chaga 2021]).

Весьма интенсивная работа по развитию корпуса была проведена в 2020–2022 гг. в рамках мегапроекта Corpus 2.0, выполняемого по гранту Министерства науки и высшего образования № 075-15-2020-793. В этот период корпус увеличился на 35 % и достиг объема, превышающего 1,5 млн слов (1308 текстов, свыше 107 тыс. предложений). Благодаря продлению гранта на 2023 год, к концу года объем корпуса еще увеличился и к началу 2024 года составил около 1 570 000 слов (1364 текста, около 111 тыс. предложений)

Корпус включает следующие типы текстов: художественная проза XX–XXI веков; современная научно-популярная литература; публицистика; биографии; новостные ленты (тексты общественно-политического, культурного, экономического и научно-технического характера).

В настоящее время СинТагРус входит в состав Национального корпуса русского языка и представляет собой отдельный его фрагмент — Синтаксический корпус (https://ruscorpora.ru/new/search-syntax.html).

Разметка СинТагРуса осуществляется в Институте проблем передачи информации РАН и по мере готовности очередного пополнения передается администраторам НКРЯ для внесения в текущую версию последнего. Информация, содержащаяся в новейшей версии Синтаксического корпуса, идентична информации в соответствующей версии СинТагРуса, однако возможности поиска по сложным запросам в Синтаксическом корпусе несколько ограничены по сравнению с поиском по СинТагРусу. Работа над расширением таких возможностей ведется, однако в силу разнообразия нетривиальной лингвистической информации в СинТагРусе, она еще не завершена. В дальнейшем изложении мы сосредоточимся именно на особенностях СинТагРуса как исходного ресурса (отмечая, впрочем, некоторые расхождения с Синтаксическим корпусом, если они являются существенными).

На данный момент в СинТагРусе имеются следующие виды разметки: 1) морфологическая; 2) синтаксическая; 3) лексико-семантическая; 4) лексико-функ-

циональная; 5) эллиптическая; 6) микросинтаксическая; 7) кореферентная; 8) темпоральная; 9) метатекстовая.

Виды разметки 1–6 покрывают весь корпус. Работа над кореферентной и темпоральной разметкой была начата относительно недавно, и в настоящее время эти типы разметки выполнены для части текстов.

Разметка текстов СинТагРуса выполняется в полуавтоматическом режиме. Сначала исходный текст обрабатывается парсером многофункционального лингвистического процессора ЭТАП-4, созданного в Лаборатории компьютерной лингвистики ИППИ РАН, в результате чего для каждого предложения строится его морфологическая и синтаксическая структура, а также идентифицируются лексические функции. Затем эти данные проверяются и корректируются экспертами-лингвистами с помощью программного комплекса «Редактор структур», разработанного специально для работы с корпусом[2]. На следующем шаге к построенным структурам применяются специализированные модули процессора, маркирующие микросинтаксические единицы, кореферентные связи и темпоральные выражения. Эти данные также проверяются лингвистами.

Морфологическая разметка состоит в том, что каждому слову текста ставится в соответствие его основная форма, часть речи и набор морфологических характеристик (значений морфологических категорий числа, падежа, вида, наклонения, времени, лица и т. д.).

Существенно подчеркнуть, что морфологическая разметка корпуса характеризуется стопроцентно снятой неоднозначностью. Она соответствует синтаксической разметке корпуса, которая также полностью однозначна.

В целом морфологическая разметка СинТагРуса соответствует принципам и конвенциям морфологической разметки ряда корпусов НКРЯ, в том числе основного корпуса НКРЯ, однако существуют некоторые расхождения. Основные расхождения состоят в следующем.

1) В СинТагРусе вид глагола рассматривается как словоизменительная категория. Какие морфологические средства используются для образования форм совершенного и несовершенного вида глагола, не имеет значения: это могут быть суффиксальные средства (*начинать — начать*), префиксальные средства (*писать — написать*), ударение (*просыпáть — просы́пать*), смешанные средства (*становиться — стать*). В случае, если у соответствующего глагола вообще существует форма несовершенного вида, именно она выбирается в качестве лексемы. Например, словоформа *стал* в выражении *Я стал инженером* представляется лексемой СТАНОВИТЬСЯ1, в выражении *Я стал в очередь* — лексемой СТАНОВИТЬСЯ2, а в выражении *Я стал хуже слышать* — лексемой СТАТЬ1 (поскольку глагол *стать*, употребляемый в качестве вспомогательного при инфинитиве, не

[2] «Редактор структур», или Structure Editor (см., в частности, [Iomdin, Sizov 2009]) — основная программная система, разработанная в Лаборатории компьютерной лингвистики ИППИ РАН специально для построения, хранения и поддержки СинТагРуса и ряда других современных компьютерно-лингвистических ресурсов.

имеет несовершенного вида: в таких случаях именем лексемы считается форма совершенного вида)[3].

2) Категория залога в СинТагРусе имеет два значения: действительный залог и страдательный залог. Медиальный залог, как в НКРЯ, не используется. В выражении *Старик с трудом поднимается со стула* словоформа *поднимается* представлена как форма действительного залога, лексема при этом — ПОДНИМАТЬСЯ, а в выражении *Этот вес штангистом еще не поднимался* словоформа *поднимался* представлена как форма страдательного залога глагольной лексемы ПОДНИМАТЬ.

3) В СинТагРусе не используется часть речи «предикатив». В большинстве случаев формы, отражаемые в НКРЯ как принадлежащие этой части речи, представлены как краткие прилагательные (*грустно* в *Маше было грустно*) или наречия (*жаль* в *Жаль, что ты не придешь*). Сведения о предикативности единицы отражаются не в морфологической, а в синтаксической структуре: в синтаксическом дереве конструкции типа *Мне было холодно* существенно отличаются от конструкций типа *Я был холоден*[4].

4) В СинТагРусе не используются «составные» части речи типа «числительное-прилагательное», «существительное-местоимение», «прилагательное-местоимение», местоименное наречие. Порядковые слова типа *четвертый* считаются прилагательными, слова типа *я, он, кто-то, что-нибудь, себя* — существительными, слова *мой, такой, любой, чей-либо* — прилагательными, слова *там, кое-где, туда* — наречиями. (Свойства местоименности таких слов, разумеется, не утрачиваются, но используются за пределами морфологии.)

5) Словоформы *его, ее, их* в притяжательной конструкции типа *его дом* представляются лексемами существительных ОН, ОНО, ОНА, ОНИ в родительном падеже, в то время как в НКРЯ такие словоформы трактуются как местоименные прилагательные без падежа с лексемами ЕГО, ЕЕ, ИХ.

6) Словоформы типа *него, нее, ним, них* представлены особыми лексемами — существительными НЕГО1 (мужского рода), НЕГО2 (среднего рода), НЕЕ, НИХ и т. д. в нужном падеже (родительном, дательном, винительном, творительном, предложном: именительного падежа у этих существительных нет). В НКРЯ здесь используются стандартные местоименные существительные ОН, ОНО, ОНА, ОНИ.

7) Вариативные формы творительного падежа (*землей — землею*) и сравнительной степени (*веселее — веселей*) представлены одинаковыми наборами

---

[3] Естественно, такое решение может влиять на тактику поиска по СинТагРусу: пользователь, удивившись, что в корпусе не обнаруживается ни одного вхождения слова *написать,* заподозрит неладное и постарается поискать слова с лексемой несовершенного вида — *писать.*

[4] Решение отказаться от использования такой части речи было принято задолго до начала разметки СинТагРуса, в ходе разработки русского синтаксического парсера ЭТАП. Важная мотивировка этого отказа состояла в следующем: авторы ЭТАПа сочли, что в выражениях типа *Казалось полезным выучить новый маршрут* и *Было полезно выучить новый маршрут* словоформы *полезным* и *полезно* относятся к одной и той же лексеме, а первая из них может быть только прилагательным.

морфологических характеристик, однако менее стандартные из них (*землею, веселей*) получают дополнительную вспомогательную характеристику alt.

8) В СинТагРусе не используется принятая в НКРЯ характеристика «2-й винительный» для представления словоформ типа *депутаты, солдаты* и т. д. в конструкциях типа *кандидат в депутаты* или *пойти в солдаты*. Эти словоформы трактуются как существительные именительного падежа множественного числа, а их синтаксический статус дополнения отражается в синтаксической структуре[5].

9) Страдательные причастия настоящего времени, образованные с помощью суффиксов типа «-ем-» и суффиксов типа «-ющ-» и частицы «-ся» (*рассматриваемый — рассматривающийся*) получают тождественные морфологические разборы.

10) Формы «смягченной» сравнительной степени с приставкой *по-* получают специальную морфологическую характеристику «смяг» в дополнение к характеристике «срав», тогда как в НКРЯ используется единая характеристика «сравнительная 2».

Синтаксическая разметка означает, что каждому предложению приписывается синтаксическая структура в виде дерева зависимостей, узлами которого являются слова предложения, а дуги помечены именами синтаксических отношений в соответствии с моделью «Смысл ↔ Текст» И. А. Мельчука.

Всего в корпусе используется 68 различных синтаксических отношений. Полный комментированный список частей речи, морфологических категорий и характеристик, а также синтаксических отношений можно найти в инструкции к Синтаксическому корпусу [Инструкция 2023].

Узлами синтаксических структур являются слова текста. В предложении естественно выделяются графические слова: цепочки алфавитных символов, разделенные пробелами и знаками препинания (внутри цепочек допускается дефис). Чаще всего один узел синтаксической структуры соответствует одному графическому слову предложения, однако в некоторых случаях СинТагРус рассматривает несколько следующих друг за другом графических слов как одну лексическую единицу — например, сочетания *как бы то ни было, ни много ни мало, до поры до времени, один на один* (в лингвистическом процессоре ЭТАП-4 такие сочетания именуются безусловными оборотами). Каждому такому сочетанию приписывается одна словарная статья, которая считается цельным словом, невзирая на пробелы. Словарь процессора ЭТАП-4 включает более 200 таких статей, и им соответствует около 11 тыс. вхождений в СинТагРусе[6].

---

[5] Стоит добавить, что в основном корпусе НКРЯ эта характеристика применяется весьма непоследовательно.

[6] В последнее время в Синтаксический корпус была введена опция, при которой элементы безусловных оборотов СинТагРуса могут по запросу пользователя отражаться в синтаксической структуре предложения порознь и представляться независимо. При этом фрагмент синтаксической структуры, соответствующий безусловному обороту, строится заранее по специально сформулированным разработчиками СинТагРуса правилам.

Лексико-семантическая разметка означает, что для многозначных (или омонимичных) слов фиксируется их лексическое значение, реализованное в данном предложении. Это значение обычно представляется именем лексемы, состоящим из леммы и следующей за ней цифры. Например, форма *толковали* в предложении *Два старика сидели за столом, пили чай и толковали о жизни* (Ю. Домбровский) представлена лексемой ТОЛКОВАТЬ1 ‘беседовать’, а такая же форма в предложении *Органы власти по-разному толковали понятие равноправия* — лексемой ТОЛКОВАТЬ2 ‘интерпретировать’. Имена лексем соответствуют именам словарных статей комбинаторного словаря лингвистического процессора ЭТАП-4. В «Редакторе структур», обслуживающем СинТагРус, предусмотрена удобная возможность заглянуть в соответствующие статьи комбинаторного словаря.

Хотя этот словарь достаточно объёмен (он насчитывает свыше 110 тысяч вхождений), разумеется, не все встречающиеся в СинТагРусе слова в нем представлены. В случае отсутствия слова в словаре ЭТАП-4 лексико-семантическая разметка ограничивается леммой. Значительную часть таких слов составляют сложные слова типа *деревообработка, Минобрнауки* или *двадцатиэтажный*. Несмотря на то, что подобные единицы вполне адекватно обрабатываются блоком композитной морфологии лингвистического процессора, в СинТагРусе было принято решение не ставить им в соответствие конкретных слов или групп слов комбинаторного словаря.

Особо следует отметить случаи, когда в корпусе присутствует слово, значение которого не отражено в комбинаторном словаре, хотя другие лексические значения этого слова в нем представлены (и, в случае неоднозначности, помечены цифрами при леммах). В таких ситуациях слово не получает никакого цифрового индекса и трактуется как не имеющее отражения в комбинаторном словаре. Заметная доля таких ситуаций приходится на случаи, когда в СинТагРусе некоторое слово представляет название, совпадающее с именем реального слова, но не вполне соответствующее ему по набору морфологических характеристик. Так, названия изданий «Коммерсант» или «Наш современник», кораблей «Викинг» или «Варяг» используют реальные слова русского языка, но выступают, в отличие от последних, как неодушевленные существительные. С другой стороны, личное имя *Урал* выступает как одушевленное существительное, в отличие от присутствующего в словаре слова *Урал* ‘горный массив’. По этой причине такие лексические единицы не наследуют свойств исходных слов и считаются отсутствующими в комбинаторном словаре.

Лексико-функциональная разметка состоит в том, что в тексте маркируются пары (а в особых случаях и тройки) слов, допускающие интерпретацию в терминах лексических функций (ЛФ) лингвистической теории «Смысл ↔ Текст». Эту разметку можно представлять себе как установление направленных связей: лексико-функциональное отношение связывает слово-аргумент лексической функции (главный член отношения) со словом, представляющим значение этой лексической функции (зависимый член отношения). В общем случае направление лексико-

функциональной связи между двумя словами не обязано совпадать с направлением синтаксического отношения между этими словами или даже предполагать наличие такого отношения.

Например, в предложении *Хлынул проливной дождь* слово *хлынул* представляет значение ЛФ INCEPFUNC0 от аргумента *дождь*, а слово *проливной* представляет значение ЛФ MAGN от *дождь*. В древовидной синтаксической структуре слово *дождь* синтаксически зависит от слова *хлынул*, тогда как лексико-функциональное отношение INCEPFUNC0 идет в обратном направлении — от аргумента *дождь* к значению *хлынул*. В то же время в этой синтаксической структуре слово *дождь* синтаксически подчиняет слово *проливной,* но и лексико-функциональное отношение MAGN имеет то же направление — от аргумента *дождь* к значению *проливной*. Если же рассмотреть предложение *Дождь был проливной*, то направление лексико-функционального отношения MAGN между *дождь* и *проливной* сохранится, а непосредственной синтаксической связи между этими словами не будет вовсе.

Эллиптическая разметка вводит в синтаксическую структуру предложения СинТагРуса слова, не представленные в графической записи предложения. Эта разметка затрагивает далеко не все типы и не все случаи эллипсиса и других разновидностей синтаксического отсутствия, существующие в языке и рассматриваемые в специальных исследованиях, и применяется достаточно избирательно. В первую очередь восстанавливаются такие типы эллипсиса, которые естественно нормализуют синтаксическую структуру присутствующей в СинТагРусе фразы (например, при опущении одного из предикатных слов в сочинительной цепочке).

В частности, при представлении предложения *Он пошел налево, а она направо* разумно принять, что между *она* и *направо* содержится лексическая единица, соответствующая слову *пошла*, но физически отсутствующая в тексте. В результате введения такого слова в синтаксическую структуру предложения оказывается естественным связать слово *пошла* с подлежащим *она* и с направительным словом *налево*. С другой стороны, в предложениях типа *Я предпочитаю светлое пиво, а мой друг любит темное* вводить для нормализации синтаксической структуры опущенное слово *пиво* в конец предложения нет необходимости, поскольку слово *темное* достаточно для того, чтобы отразить наличие у глагола *любит* прямого дополнения. Подробнее об эллиптической разметке см. ниже в разделе 2.

Микросинтаксическая разметка идентифицирует синтаксически своеобразные фразеологические и полуфразеологические выражения самых разных типов (именуемые микросинтаксическими единицами), содержащиеся в текстах СинТагРуса и отличающиеся своеобразием синтаксического поведения. Примерами таких единиц могут быть обороты типа *всё равно* (в разных значениях), составные союзы типа *как будто бы, разве что,* парные союзы типа *не только ... но и, не то ... не то,*

составные предлоги типа *со стороны, в качестве,* адвербиалы типа *с виду, на виду, про запас* и многие другие. Число разных микросинтаксических единиц, которые встречаются в корпусе СинТагРус, превышает 3200.

Этот тип разметки применяется относительно недавно, однако в настоящее время уровень покрытия достиг всего объема СинТагРуса, а текущая разметка осуществляется одновременно с пополнением корпуса новыми текстами. Подробнее микросинтаксическая разметка будет рассмотрена в разделе 3.

Кореферентная разметка предполагает установление особых связей между словами текста, референты которых совпадают, включая анафорические связи между местоимениями и их антецедентами. В отличие от охарактеризованных выше типов разметки, кореферентная разметка не ограничивается отдельными предложениями, а покрывает весь текст целиком. Благодаря этому обстоятельству кореферентные цепочки в одном тексте могут быть весьма длинными и достигать сотен элементов.

Создатели СинТагРуса приступили к кореферентной разметке относительно недавно, причем в первом варианте разметка сводилась только к установлению местоименной анафоры и выполнялась в экспериментальном режиме для текстов корпуса, размеченных в 2017–2019 гг. [Иншакова и др. 2019]. С развитием проекта Corpus 2.0 и разработкой механизма установления неместоименной кореферентности было принято решение существенно увеличить объем подлежащих кореферентной разметке текстов. В соответствии с этим решением объем текстов СинТагРуса с кореферентной разметкой к концу 2023 года составил более 120 тыс. слов, или около 9200 предложений.

Данный тип разметки будет подробнее рассмотрен в разделе 4.

Темпоральная разметка предназначается для идентификации в тексте выражений, описывающих протекание событий во времени, а также связей между ними. Принципы темпоральной разметки СинТагРуса были подробно описаны в работе [Тимошенко и др. 2021]. В 2020–2022 годах темпоральная разметка СинТагРуса была выполнена для нескольких текстов корпуса. Однако в текущей версии СинТагРуса развитие этого типа разметки и пополнение ею новых текстов не производится; к ней предполагается вернуться через некоторое время.

В рамках проекта Corpus 2.0 была проведена большая работа по совершенствованию **метаразметки** СинТагРуса. Формат метатекстовой разметки был приведен в соответствие со стандартом, принятым в основном корпусе НКРЯ и описанным в [Савчук 2005]. Исключением являются, в частности, поля «возраст аудитории» и «размер аудитории», которые не были использованы нами из-за неоднозначности их трактовки и их непоследовательного использования в НКРЯ. Поля «хронотоп» и «тематика текста» в СинТагРусе объединены в одно поле, поскольку первое указывается только для художественных текстов, а второе — только для нехудожественных. Кроме того, есть некоторые поля, присутствующие в СинТагРусе, но отсутствующие в НКРЯ, включая год разметки текста, часть текста

(если размечен фрагмент), имена аннотаторов текста и дополнительные комментарии.

Общее число полей метатекстовой разметки возросло с 7 до 17. Заполнение новых полей и правка старых для всех текстов корпуса были проведены вручную с использованием «Редактора структур», в который был добавлен функционал, обеспечивающий единообразие разметки между разными текстами и частями одного текста. В том числе обеспечивается одинаковость написания имени одного и того же автора в разных текстах.

На рис. 1 приведен пример метаразметки текста СинТагРуса в варианте Синтаксического корпуса НКРЯ. Отличие от разметки в самом СинТагРусе состоит в отсутствии некоторых дополнительных полей и включении пола и года рождения автора (в СинТагРусе эта информация является частью сведений об авторе, хранящихся отдельно от самих текстов).

Автор	Михаил Афанасьевич Булгаков
Пол	мужской
Год рождения	1891
Название	Мастер и Маргарита
Год создания	1928-1940
Сфера функционирования	художественная
Жанр текста	фантастика
Тематика текста	ирреальный мир
Тип текста	роман
Носитель	электронный текст
Источник	http://bibliotekar.ru/bulgakov/2.htm
Дата публикации	-
Издание	Библиотекарь.Ру
Издательство	-
Часть	глава 2
Год разметки	2018

Рис. 1. Метаразметка текста 2-й главы «Мастера и Маргариты» (в варианте Синтаксического корпуса НКРЯ)

2. Несколько иллюстративных примеров

Для иллюстрации основных видов разметки СинТагРуса (морфологической, синтаксической, лексико-семантической, лексико-функциональной и эллиптической) мы рассмотрим подробнее представление нескольких реальных примеров предложений, содержащихся в корпусе.

Для начала посмотрим, как в СинТагРусе отражено предложение

(1) *Тут прокуратор поднялся с кресла, сжал голову руками, и на желтоватом его бритом лице выразился ужас.* (М. Булгаков, «Мастер и Маргарита»).

В «Редакторе структур» предложение (1) выглядит следующим образом:

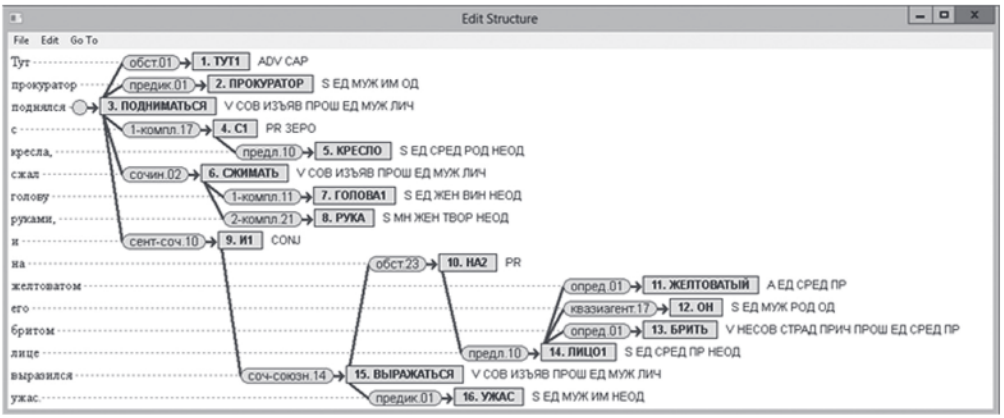

Рис. 2. Структура предложения (1)

Для каждого слова предложения, записываемого слева, в прямоугольной рамке указано соответствующее ему имя лексемы, связывающее его со статьей комбинаторного словаря лингвистического процессора ЭТАП-4, и представляющее собой лемму с добавленным цифровым индексом в случае неоднозначности (все равно, омонимии или полисемии). Например, ТУТ1 — статья наречия (а не существительного), С1 — предлог, управляющий родительным падежом, ГОЛОВА1 — часть тела (а не начальник), И1 — союз (а не частица), НА2 — предлог, управляющий предложным падежом; ЛИЦО1 — неодушевленное существительное со значением «часть тела», а не одушевленное существительное со значением «персона». Это и есть лексико-семантическая разметка корпуса.

Если разметчики СинТагРуса или его пользователи захотят удостовериться в том, что в разметке выбрано нужное значение слова, они могут воспользоваться особой опцией «Редактора структур», чтобы в реальном времени заглянуть в комбинаторный словарь и увидеть, каким лексическим значением представлено соответствующее слово. В частности, воспользовавшись данной опцией, разметчик или пользователь сможет увидеть представительные фрагменты словарных статей слов, основные формы которых совпадают, и подтвердить или изменить сделанный в СинТагРусе выбор.

Так, для неоднозначного слова ТУТ «Редактор структур» покажет по запросу два фрагмента комбинаторного словаря: для наречия ТУТ1 (рис. 3) и для существительного ТУТ2 ‘тутовое дерево’ (рис. 4).

Аналогичным образом, при просмотре лексических значений слова ГОЛОВА мы сможем увидеть фрагменты словарных статей, соответствующих лексеме ГОЛОВА1 (‘часть тела’) и ГОЛОВА2 (‘начальник’, как в *городской голова*). При этом демонстрация разных лексических значений слова не зависит от того, насколько реально появление этих значений в рассматриваемой фразе: система показывает все значения.

Рис. 3. Фрагмент словарной статьи ТУТ1, доступный при работе с СинТагРусом над предложением, содержащем слово ТУТ

Рис. 4. Фрагмент словарной статьи ТУТ2, доступный при работе с СинТагРусом над предложением, содержащим слово ТУТ

Справа от каждой прямоугольной рамки на рис. 2 указывается часть речи и список морфологических характеристик слова: напомним, что как морфологическая, так и синтаксическая структуры в СинТагРусе являются стопроцентно дизамбигуированными.

Основные морфологические характеристики, представленные на рис. 2, кажутся самоочевидными.

В овалах записаны стандартно сокращенные имена синтаксических отношений (обстоятельственное, предикативное, 1-е комплетивное и т. п.). Как нетрудно увидеть, при каждом слове записано имя единственного синтаксического отношения, входящего в это слово; существует, однако, ровно одно слово (вершина дерева зависимостей), в которое не входит никакое синтаксическое отношение. В примере (1) такой вершиной является третье по порядку слово — глагол *поднялся.*

Цифры в овалах после имени синтаксического отношения указывают на номер синтаксического правила (синтагмы), пользуясь которым синтаксический парсер процессора ЭТАП-4 установил данное отношение. Данный тип информации пред-

назначается для разработчиков СинТагРуса, а не для его пользователей. В Синтаксическом корпусе эти номера не воспроизводятся.

Рассмотрим еще одно предложение СинТагРуса, чтобы подробнее проиллюстрировать его лексико-функциональную разметку:

(2) *Присуждение Нобелевской премии этого года в области физиологии и медицины прошло под знаком точных наук*. (Ф. Смирнов, «Нобелевские премии: физики снова в почете»)

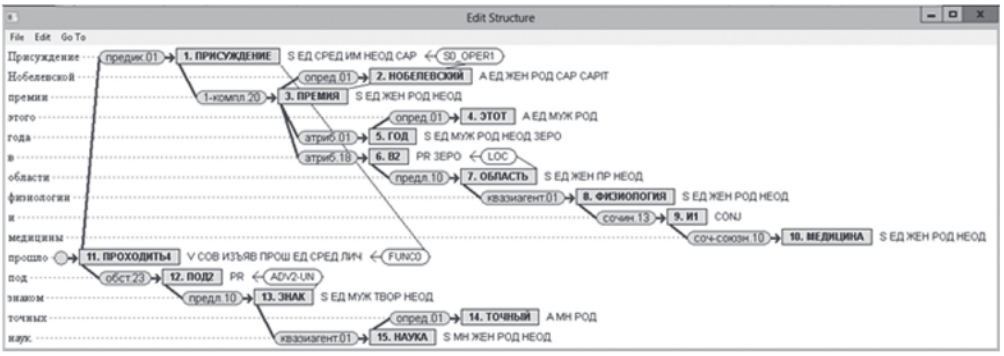

Рис. 5. Структура предложения (2)

Как уже отмечалось, лексико-функциональная разметка фиксирует пары слов в предложении, между которыми можно установить связи в терминах лексических функций модели «Смысл ↔ Текст». В предложении (2) таких пар слов (лексических коррелятов) четыре: (а) *присуждение премии*, (б) *прошло присуждение*; (в) *в области* и (г) *под знаком*. Во всех четырех парах первое слово является значением некоторой лексической функции, а второе — аргументом (ключевым словом) этой функции. В паре (а) имеет место ЛФ S0_OPER1; это по существу субстантивация ЛФ OPER1 (‘делать то, что обозначено ключевым словом’) — *присуждать премию* → *присуждение премии*, в (б) — ЛФ FUNC0 (‘иметь место’), в (в) — ЛФ LOC (типичный предлог, характеризующий ключевое слово в качестве места) и в (г) — ЛФ ADV2-UN (типичный предлог, образующий адвербиал, который характеризует ситуацию, описываемую ключевым словом, с точки зрения его второго участника). Во всех четырех случаях аргумент и значение ЛФ синтаксически связаны между собой (что, как мы видели в разделе 1, бывает не всегда), причем в трех случаях (а, в и г) связь направлена от значения ЛФ к аргументу, а в одном случае (б) — от аргумента к значению.

На рис. 5 направление лексико-функциональной связи графически представлено стрелками, внешний вид которых отличается от представления синтаксических отношений.

Добавим в заключение, что лексико-функциональные связи в СинТагРусе представлены одним из двух широких классов ЛФ — так называемыми ЛФ-параметрами. Другой широкий класс ЛФ — ЛФ-замены, такие как ЛФ SYN (сино-

ним), ANTI (антоним), CONV (конверсив) и ряд других (в том числе синтаксические дериваты), в корпусе не отражается. Такое решение было принято потому, что, как правило, в тексте практически не встречаются одновременно аргументы и значения таких функций. Между тем в некоторых ситуациях значения ЛФ данного класса все-таки появляются в предложении вместе с их аргументами; например, в случае ЛФ GENER (родовое слово). Помимо хорошо известных цитат типа *Спит животное Собака, Дремлет птица Воробей* (Н. Заболоцкий), где слова *животное* и *птица* выступают как значения ЛФ GENER при словах *собака и воробей* соответственно, аналогичные примеры встречаются и во вполне нейтральных текстах, в том числе и в СинТагРусе; ср., например

(3) *Появился новый препарат эпоэтин, близкий к гормону эритропоэтину, способному повышать выработку эритроцитов и тем самым улучшать снабжение организма кислородом и справляться с такими видами малокровия, которые ранее были неизлечимы* (В. Прозоровский, «Лекарства от усталости»),

где слова *препарат и гормон* выступают в качестве значения ЛФ GENER к словам *эпоэтин* и *эритропоэтин* соответственно. Сейчас такие ЛФ не фиксируются, однако обсуждается возможность дополнить лексико-функциональную разметку и некоторыми функциями, относящимися к классу замен.

Комментированный список всех используемых в СинТагРусе лексических функций приводится в инструкции к Синтаксическому корпусу [Инструкция 2023]. В настоящее время в текстах корпуса встречается 143 различных лексических функции, они реализованы приблизительно в 44 тыс. пар слов. Количество предложений, содержащих хотя бы одно вхождение ЛФ, составляет около 32 тыс., или 30% всех предложений корпуса.

Обратимся теперь к одному из предложений СинТагРуса, содержащих эллиптическую разметку:

(4) *— Взять бы этого Канта, да за такие доказательства года на три в Соловки! — совершенно неожиданно бухнул Иван Николаевич* (М. Булгаков, «Мастер и Маргарита»).

Представляется очевидным, что в предложении (4) опущено некоторое глагольное слово, иначе весьма затруднительно понять, к чему могут относиться предложно-именные сочетания *за такие доказательства, года на три* и *в Соловки.* Разработчики СинТагРуса приняли решение в подобных случаях восстанавливать такие опущенные предикаты, как и прототипический эллипсис, помещая в определенное место фразы соответствующую лексическую единицу. Технически такое решение осуществляется следующим образом: (1) выбирается конкретная позиция в предложении; (2) на эту позицию помещается опущенное слово, которому приписывается определенное словарное имя, нулевая словоформа и нужный набор морфологических характеристик; (3) это опущенное слово встраивается в синтаксическую структуру предложения. Само опущенное слово получает поряд-

ковый номер в предложении и дополнительную помету Phantom. Результат такой обработки предложения (4) представлен на рис. 6:

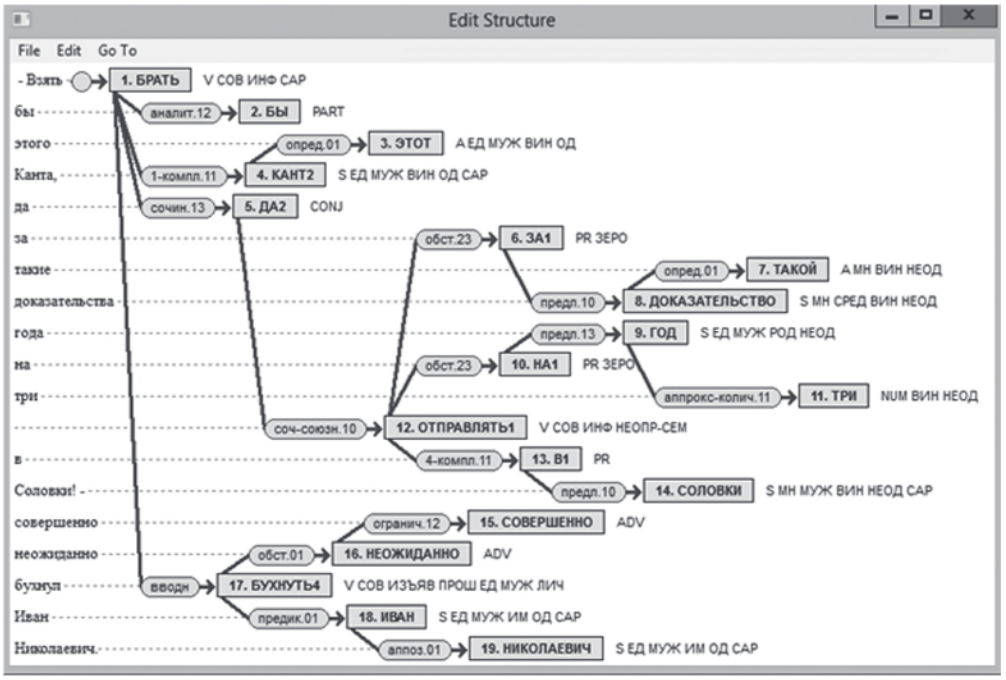

Рис. 6. Структура предложения (4)

Как и в случаях со стандартными узлами структуры, информацию о лексических свойствах элидированного слова можно получить посредством описанной выше опции обращения к комбинаторному словарю (см. рис. 7).

Содержание словарной статьи ОТПРАВЛЯТЬ1 можно посмотреть, нажав на кнопку «. . .» справа от поля «КС Name».

Поскольку лингвистический процессор ЭТАП-4 не располагает механизмом автоматического восстановления эллипсиса, узлы с пустыми текстовыми элементами («фантомы») добавляются в синтаксическую структуру предложения вручную в ходе ее проверки и коррекции лингвистами (после работы этапа автоматического синтаксического анализа, конструирующего морфосинтаксическую структуру предложения).

В большинстве случаев добавление «фантомов» в структуру не вызывает затруднений, так как они повторяют другие лексические единицы текста (разумеется, с возможной морфологической модификацией). Если же параллелей в окружающем тексте нет, ситуация становится более неопределенной. Именно так обстоит дело в примере (4): подходящим кандидатом в «фантомы» будут многие глаголы: *отправить, выслать, депортировать, загнать* и др. (причем, учитывая личность персонажа текста, глагол *отправить* окажется чересчур мягким). Тем не менее в подобных случаях лингвист, осуществляющий разметку, выбирает по своему

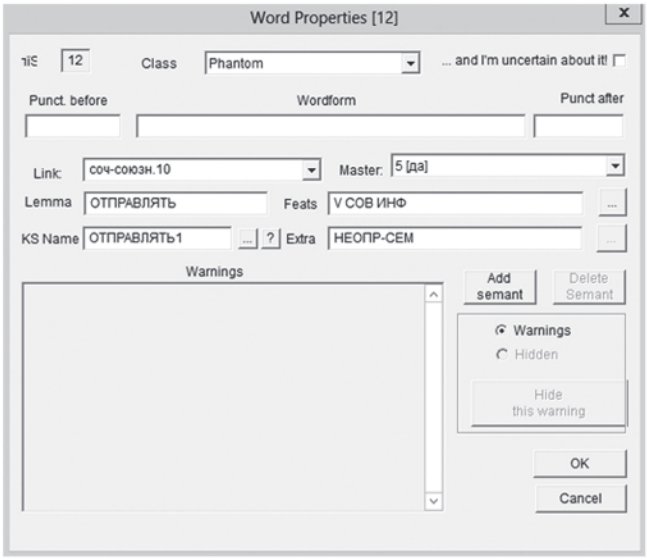

Рис. 7. Карточка узла фантомной словоформы *отправить*, доступная при работе с СинТагРусом над предложением, содержащем слово ОТПРАВЛЯТЬ1 с признаком Phantom

усмотрению один из семантически наиболее нейтральных вариантов и приписывает новому узлу специальный признак «неопр-сем» (неопределенная семантика).

Аналогичным образом разметчики поступают и в других подобных ситуациях. Скажем, в предложении

(5) *Татьяна в лес; медведь за нею...* (А. Пушкин, «Евгений Онегин»)

подходящими кандидатами в «фантомы» будут многие глаголы, означающие быстрое движение: *Татьяна ... бросилась, помчалась, побежала...; медведь ... погнался, побежал...*

В настоящее время в корпусе насчитывается около 4 тыс. «фантомов» эллиптической разметки, в том числе 450 с признаком «неопр-сем».

В отдельных случаях при эллиптической разметке корпуса оказывается затруднительным не только зафиксировать конкретное опущенное слово, но даже сколько-нибудь надежно определить семантический класс таких слов. В подобных ситуациях мы прибегаем к лексическому заполнению «фантома» символом неопознанного слова, по возможности фиксируя полный или частичный набор его морфологических характеристик. Примерами могут служить незаконченные предложения СинТагРуса типа

(6) *Поскольку изоляция большой страны с ядерным оружием...*

или

(7) *Между прочим, здоровые люди в это время пьют водку, а я...*

В предложение (6) после слова *оружием* добавляется «фантом», обозначаемый как фиктивный глагол настоящего времени; он синтаксически подчинен подчинительному союзу *поскольку*, а слово *изоляция* подчинено этому «фантому», выступая в роли подлежащего при нем. В предложении (7) «фантом» также помещается в конец предложения и также обозначается как фиктивный глагол настоящего времени. Он синтаксически подчинен сочинительному союзу *а* и, в свою очередь, подчиняет слово *я* в качестве подлежащего.

3. Микросинтаксическая разметка СинТагРуса

Микросинтаксическая разметка СинТагРуса осуществляется разработчиками корпуса начиная с 2016 года. Решение пополнить корпус этой разметкой было обусловлено появлением теории микросинтаксиса (см., в частности, [Иомдин 2008] и [Iomdin 2016]), предлагающей принципы описания широкого круга языковых явлений, лежащих на границе словаря и грамматики. Эта теория имеет много общего с Грамматикой конструкций Ч. Филлмора и его коллег (начиная с ранних работ, таких как [Fillmore et al. 1988]), однако развивалась совершенно независимо от Грамматики конструкций, первоначально на материале русского языка. Следует также добавить, что теория микросинтаксиса и создаваемые на ее основе ресурсы (микросинтаксическая разметка СинТагРуса, Микросинтаксический словарь русского языка) имеют ряд общих черт с разрабатываемым с недавнего времени лингвистическим ресурсом «Русский конструктикон» [Endresen et al. 2020], однако и в данном случае соответствующие подходы и ресурсы развиваются независимо.

Основные принципы и подходы к микросинтаксической разметке СинТагРуса были подробно изложены в [Иншакова и др. 2019] и здесь не повторяются. Мы ограничимся здесь сведениями о современном состоянии этой разметки и о возможностях ее использования в лингвистических исследованиях.

В настоящее время микросинтаксическая разметка распространяется на весь СинТагРус целиком. Число разных микросинтаксических единиц, отраженных в разметке СинТагРуса, составляет 3220. Меньшую часть этих единиц (не более ста) представляют нестандартные синтаксические конструкции (главным свойством которых является синтаксическая специфика и минимальное число конкретных лексических единиц).

Это, например, сформированные дательным падежом и инфинитивом инфинитивно-модальные конструкции — с отрицанием; ср.

(8) *Ведь* ***никому не понять****, что с ней происходит и почему она в троллейбусе* (Ю. Трифонов) ≈ ‘Отсутствует перспектива, что кто-нибудь сможет понять...’

или без него; ср.

(9) ***Мне*** *еще на поезд* ***успеть****...* (В. Шукшин) ≈ ‘Я должен успеть на поезд’.

Среди инфинитивно-модальных конструкций особо выделяются конструкции с вопросительными словами, причем вопросительные слова могут быть за преде-

лами самой конструкции, как в (10) или (12), или занимать позицию дательного субъекта в ней, как в (11):

(10) — *А* ***для чего тебе владеть*** *звездами?* (А. де Сент-Экзюпери, «Маленький принц»);
(11) ***Какому быть*** *главнее*? (М. Шиманский, «Василь Быков: язык — душа народа»);
(12) — *Проблема двуязычия легче разрешима, скажем, в Латвии или Армении, а* ***как быть уйгуру или дунганину*** *в Киргизии?* (Ibid.)[7]

Другой подкласс нестандартных синтаксических конструкций представлен единицами, содержащими два вопросительных местоимения. Главным компонентом значения этой конструкции является распределение элементов двух множественных объектов, участвующих в ситуации; ср.

(13) *Важны были другие знания:* ***кто что*** *спрашивает,* ***кому как*** *отвечать (один любит сразу, другой — подумавши), как легче заучить наизусть формулы или формулировку* (И. Грекова, «Кафедра»);
(14) *Можно отказаться от обсуждения проблем: мало ли* ***что где*** *происходит — кому надо, те разберутся* (Г. Онищенко, «Как закупать дорогостоящее оборудование»).

Целый ряд нестандартных синтаксических конструкций представляют собой единицы с лексическими повторами и с весьма нетривиальным значением; например, $X_{пов}$ *не* $X_{пов}$:

(15) ***Работай не работай****, а денежки идут* (Л. Радзиховский, «Вызов и ответ»);

$X_{пов}$ *да* $X_{пов}$:

(16) *... это все Пашка, пристал ко мне:* ***напиши да напиши*** (А. Мамедов, «Дзэн в городе N»).

В некоторых случая лексические повторы требуют присутствия в контексте других конкретных слов. Таково, например, слово ***рознь*** в единице *X X-у рознь:*

(17) *Ошибки практически неизбежно сопровождают любое научное исследование, но* ***ошибки ошибкам рознь*** (Е. Александров, Ю. Ефремов, «Как воевать с лженаукой»).

---

[7] Не вдаваясь в сложные подробности, отметим, что предложение (12) представляет собой пример весьма частой ситуации, когда две микросинтаксические единицы частично пересекаются: в данном случае это инфинитивно-модальная конструкция рассмотренного только что типа и синтаксическая фразема *как быть.* Аналогичная ситуация имеет место и в предложении (14), где частично пересекается синтаксическая фразема «мало ли + ВОПР» и нестандартная синтаксическая конструкция с двумя вопросительными словами.

Подавляющее большинство микросинтаксических единиц, размеченных в СинТагРусе, относится к классу синтаксических фразем. Среди разрядов таких единиц можно выделить, в частности, 1) отдельные неоднословные выражения, более или менее эквивалентные словам, в том числе многозначные, типа *все равно* (в некоторых значениях), *через раз, на ночь глядя* и десятки других; 2) составные союзы и предлоги типа *как только, потому что, пока что, тогда как; со стороны, в качестве, по поводу,* 3) предложно-именные слабокомпозициональные или некомпозициональные словосочетания типа *с виду, на вид, на виду, под видом, в бытность, в возрасте, в миру, под рукой, в духе* (в разных значениях: (а) ‘на принципах’, как в *воспитывать в духе толерантности* и (б) ‘в определенном настроении’, как в *Он сегодня совершенно не в духе*) и т. д.; 4) некомпозициональные словосочетания, принципиально не сводимые к единому слову, такие как *то и дело, то ли дело, в том числе, в упор не видеть* и сотни других).

Микросинтаксическая разметка корпуса осуществляется в «Редакторе структур». На рис. 8 показано, как выглядит эта разметка для фрагмента текста корпуса, а на рис. 9 представлена разметка одного предложения.

В левой части окна, представленного на рис. 8, помещены отдельные предложения текста, а в правой его части отмечены микросинтаксические единицы, встречающиеся в соответствующих предложениях. Например, единица *по тем временам* присутствует в предложении 3 и занимает отрезок от слова 2 до слова 4, а единицы *как только* в предложении 21 и *в одиночку* в предложении 26 занимают

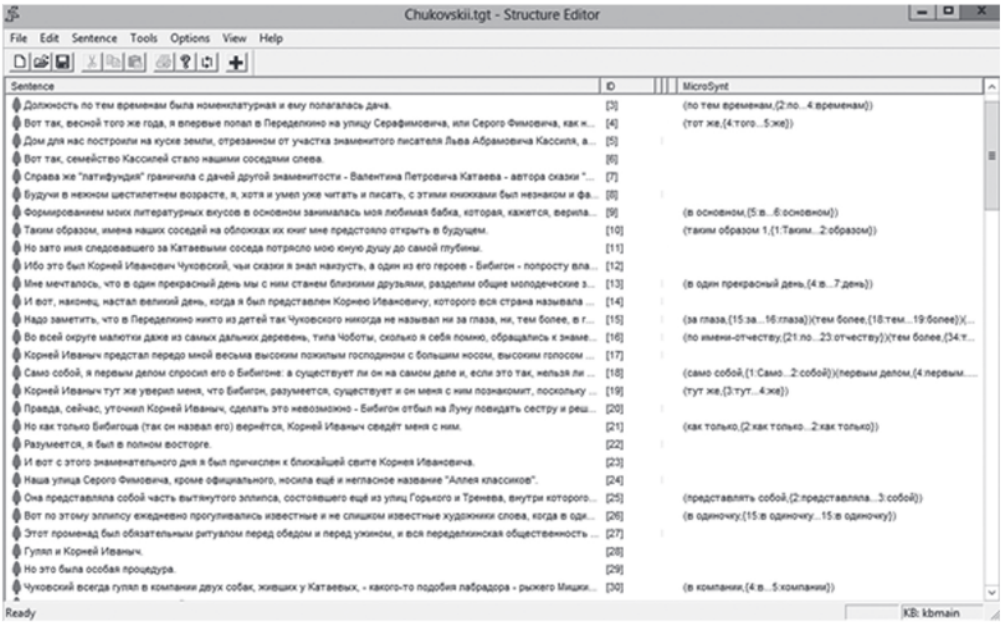

Рис. 8. Микросинтаксическая разметка фрагмента эссе К. Смирнова «Улица Серого Фимовича»

отрезок длиной в одно-единственное слово (2 в предложении 21 и 15 в предложении 26). Последние два случая объясняются тем, что в число микросинтаксических единиц входит большинство безусловных оборотов, которые лингвистический процессор ЭТАП-4 трактует как единые слова.

В нижней части рисунка перечисляются присутствующие в предложении четыре микросинтаксические единицы: составные союзы *да и* и *так что* (последний занимает отрезок в одно слово), временная частица *только что* и местоименное прилагательное тождества *тот же.*

Мы уже отмечали, что микросинтаксическая разметка осуществляется на пространстве всего корпуса СинТагРус. Результаты этой разметки, осуществляемой почти исключительно вручную, можно резюмировать следующими цифрами.

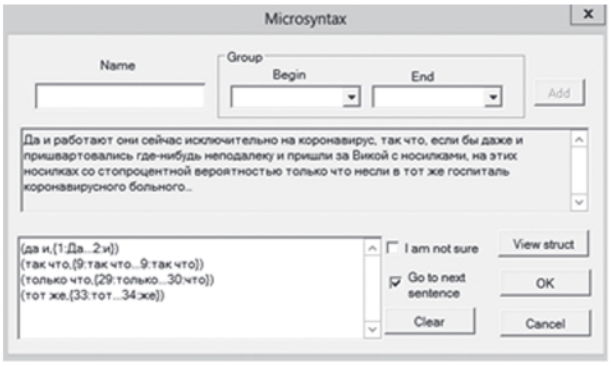

Рис. 9. Микросинтаксическая разметка предложения из очерка Н. Раппопорт «Набережная исцелимых»

Число микросинтаксических единиц, входящих в микросинтаксическую разметку, составляет более 45 600. При этом число предложений, содержащих хотя бы одну микросинтаксическую единицу, составляет около 34 300. Разница между двумя последними цифрами естественно объясняется тем, что многие предложения содержат по несколько микросинтаксических единиц. Доля предложений СинТагРуса, в которых присутствует хотя бы один микросинтаксический элемент, составляет почти 32 %. Этот факт, на наш взгляд, убедительно свидетельствует о высокой степени базовой (нефразеологической) идиоматичности русского текста.

Важно при этом подчеркнуть следующее. Несмотря на то, что объем микросинтаксической разметки СинТагРуса достаточно велик (насколько известно авторам, в мире не существует корпусов текстов со сравнимым количеством детектированных и тщательно атрибутированных в тексте идиоматичных выражений), его все еще недостаточно для полномасштабных исследований функционирования таких единиц в языке. В частности, многие микросинтаксические единицы присутствуют в корпусе в весьма ограниченном количестве: из 3220 таких единиц 1235 (почти 40 %) встречаются в текстах корпуса всего один раз, в то время как количество микросинтаксических единиц, присутствующих в корпусе 50 раз и более, составляет всего 200.

Кроме того, ввиду ограниченности объема всего СинТагРуса, определенные микросинтаксические единицы, в том числе весьма интересные как с синтаксической, так и семантической точек зрения, вообще туда не попали (таковы, например, единицы *под шумок, дать дёру* или *вдруг да и: Бывают в жизни тупики, говорила я, которые только кажутся тупиками, а* ***вдруг да*** *и расступятся.* (Л. Чуковская, «Предсмертие»)[8].

Тем не менее, несмотря на недостаточную представительность микросинтаксической разметки корпуса, даже при этом объеме удается провести ряд интересных лингвистических исследований. Завершая раздел о микросинтаксической разметке, мы кратко изложим в качестве примера результаты одного из таких исследований.

При анализе материалов микросинтаксической разметки СинТагРуса обращает на себя внимание следующий факт: среди микросинтаксических единиц обнаруживается неожиданно много выражений, сформированных существительными во втором родительном (будем для краткости именовать его партитивным, независимо от того, передает он значение партитивности или нет) и во втором предложном (местном) падежах. В первую очередь это единицы, состоящие из предлога и управляемого им существительного и в целом синтаксически более или менее эквивалентные наречию (приглагольному или прииименному).

Единицы с партитивным падежом включают, например, такие выражения, как *с виду, с ходу, с лёту, с разбегу, с налёту, с маху, с размаху, с краю, с краешку, с перепугу, с перепою, с глузду, без толку, без спросу, без умолку, без удержу, из дому, от роду, от веку, до упаду, до зарезу, до свету* ‘до рассвета’.

Легко заметить, что многие существительные, образующие эти выражения, за пределами микросинтаксических единиц не встречаются или почти не встречаются: существительные *перепуг, перепой, глузд, удерж, умолк, упад, зарез* сами по себе в современных текстах не фигурируют.

Кроме того, в число конструкций с партитивом входят и другие единицы, с обязательным или факультативным участием других слов (в частности, с прилагательным — определением к существительному в партитиве, ср. *со всего маху <размаху>, с какого боку, с одного боку, с другого боку, без всякого толку*, разнообразные конструкции с повторением либо существительного, либо предлога (*с глазу на глаз, с боку на бок, час от часу, с пылу с жару*) и отдельные полилексемные конструкции с предлогом (*баба с возу, без году неделя, без роду и племени, с миру по нитке, с бору по сосенке, сбивать с толку, сбивать с панталыку, терять из виду, не до жиру, беситься с жиру, с боку припёка*) или без него (*что толку, толку чуть, спору нет, сносу нет, нашего полку прибыло*) и т. д.

Доля таких единиц в общем объеме выражений современного русского языка, содержащих партитив, непропорционально высока. В частности, при поиске в основном корпусе НКРЯ выражений, содержащих ***S, gen2***, из 100 первых фрагментов текстов, в общей сложности содержащих 159 вхождений партитива, нефразео-

[8] Следует, впрочем, оговориться, что в процессе многолетней работы над микросинтаксическими единицами в поле зрения авторов оказалось весьма небольшое количество таких единиц, не представленных в корпусе.

логических сочетаний с существительным в партитиве обнаружилось 29 (18 %); во всех остальных случаях мы имеем дело с бесспорными фразеологическими выражениями (почти исключительно с микросинтаксическими конструкциями).

Еще красноречивее оказывается распределение фразеологических и нефразеологических выражений среди сочетаний «предлог + партитив»: 100 первых фрагментов текстов основного корпусе НКРЯ, удовлетворяющих запросу ***Pr+gen2***, содержат 248 таких сочетаний, из которых только 8 (3,2 %) являются нефразеологическими (*вместо табаку, [поднял] с полу, [умер] с голоду, после чаю* и некоторые другие).

Единицы со вторым предложным падежом представлены, в частности, выражениями *на ходу, на бегу, на лету, в ходу, в строю, в краю* и т. д. В этих конструкциях вырисовывается похожая картина: доля микросинтаксических единиц среди таких конструкций тоже весьма высока, хотя и не столь впечатляюща (в процентном отношении их несколько меньше).

Так, в первых ста фрагментах текстов, полученных по запросу к основному корпусу НКРЯ ***Pr+loc2***, встречается 191 такое словосочетание, из которых не менее 120 (63 %) следует расценивать как микросинтаксические единицы (правда, некоторые их них, например, *в виду* или *в связи с*, встречаются многократно).

По существу, каждая микросинтаксическая единица рассматриваемого здесь типа требует индивидуального исследования — даже с точки зрения взаимодействия предлога с конкретным падежом существительного. Ср., например, микросинтаксическую единицу *на дому* с весьма специфическим значением, указывающим на то, что какая-то профессиональная деятельность осуществляется в месте, где живет человек, занимающийся этой деятельностью (в качестве активного участника или клиента), как в примере

(18) *Домашние сборища запрещали, но Ляпунов продолжал читать на дому лекции по теории программирования.* (Д. Гранин, «Зубр»)

и нефразеологическое выражение *в дому* в песне А. Галича:

(19) *Я живу теперь в дому — чаша полная, Даже брюки у меня и те на «молнии».*

И в том, и в другом случае мы встречаем форму второго предложного *дому,* но *на дому* — это микросинтаксическая единица, а *в дому* — это просторечие (может быть, стилизованное).

Конечно, обнаруженные нами закономерности в употреблении второго родительного и второго предложного падежей носят вероятностный, а не абсолютный характер. Ср, например, нефразеологическое выражение *в поту* со вторым предложным, как в примере

(20) *Весь пол был заляпан белыми пятнами, все мы были в поту*. (М. Булгаков, «Полотенце с петухом»)

и фразеологическую единицу *в поте лица* в выражении типа *трудиться в поте лица.*

Тем не менее в совокупности приведенные факты убедительно свидетельствуют о консервативности фразеологических единиц, массово сохраняющих в языке уходящие формы, и о морфологической устойчивости таких единиц.

Верно и обратное утверждение: уходящие из языка реликты находят своеобразное убежище во фразеологии и тем самым обретают достаточно надежную нишу, в которой продлевают свое присутствие в языке в целом.

4. Разметка кореферентных связей в СинТагРусе

В конце 2023 г. объем размеченной такими связями части СинТагРуса составил более 120 тыс. слов, или около 9200 предложений в 33 текстах.

Особенность текстов с кореферентной разметкой из СинТагРуса заключается в том, что почти все они имеют достаточно большую длину по сравнению с текстами из других корпусов с кореферентной разметкой. На данный момент длина таких текстов в нашем корпусе составляет от 30 до 685 предложений (для сравнения, в корпусе RuCor максимальная длина текста — 170 предложений [Toldova et al. 2014], в тестовом корпусе AnCor [Budnikov et al. 2019] — 336 предложений, в RuCoCo [Dobrovolskii et al. 2022] — 202 предложения, в ARRAU (подкорпус GNOME) [Uryupina et al. 2020] — около 320 предложений, в Пражском корпусе зависимостей [Hajič et al. 2018] — 231 предложение). Это обусловлено выбором для СинТагРуса текстов определенных жанров — публицистических и научно-популярных статей, интервью, глав из художественной литературы и мемуаров (многие из них входят и в Уппсальский корпус русских текстов). Одним из следствий этого являются большие расстояния между некоторыми элементами кореферентных цепочек, которые могут составлять сотни предложений.

Кореферентные связи в СинТагРусе проводятся поверх синтаксической разметки. Поскольку она представляет собой деревья зависимостей без указания границ составляющих, кореферентные связи соединяют словоформы — вершины именных групп или некоторых других составляющих и не содержат информации об их границах (однако существует возможность вычислить эти границы по соответствующим поддеревьям с помощью разработанной в лаборатории программы).

Одна кореферентная связь соединяет два упоминания одной сущности (в направлении от правого к левому, за исключением случаев катафоры) и содержит их координаты в тексте вида **(-*y*;)*x*:**, где *x* — номер словоформы в предложении (обязательная координата), а *y* — количество предложений между первым и вторым упоминаниями (указывается, если они находятся в разных предложениях). Анафорические связи от возвратных и взаимно-возвратных местоимений проводятся к их ближайшим неместоименным антецедентам, а кореферентные связи от неместоименных существительных и местоимений 3-го лица — к ближайшим неместоименным упоминаниям той же сущности. Если антецедент является сочиненным или расщепленным, то от местоимения проводятся связи к каждому из его элементов. Из отдельных пар упоминаний с помощью специальной программы можно получить упорядоченные множества упоминаний, соответствующие сущностям. Вид

кореферентной разметки фрагмента текста СинТагРуса в окне «Редактора структур» приведен на рис. 10.

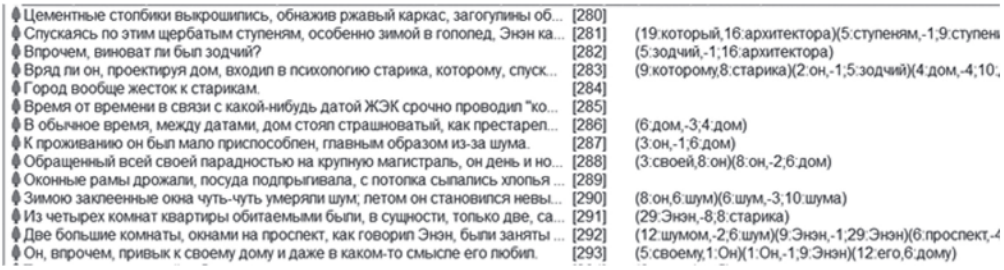

Рис. 10. Фрагмент СинТагРуса с кореферентной разметкой

В корпусе размечаются следующие виды связей:

1) анафорические связи:

- от возвратных (*себя, свой*) и взаимно-возвратных (*друг друга / друг дружку*) местоимений — в любых случаях, включая связанную анафору с кванторными антецедентами (*...никто*$_i$ *из них, даже кляня свою*$_i$ *судьбу, не поменял бы ее ни на чью другую*);
- от местоимений 3-го лица, включая «местоимение переключения референции» *тот*, — везде, за исключением случаев, когда местоимение и его антецедент отсылают к разным референтам (*Во-вторых, создать условия, вынуждающие руководителей предприятий и организаций не заказывать лишних специалистов*$_i$*, а эффективнее их*$_j$ *использовать*);
- от всех видов относительных местоимений, включая наречные (*который, кто, что, чей, какой, где, куда, откуда, когда*);
- от местоимений-наречий *здесь, тут, там, туда, оттуда, тогда*. Если их антецедентом является не существительное или наречие, а предложная группа, то кореферентная связь проводится к предлогу (*в*$_i$ *доме / у*$_i$ *дома / за*$_i$ *домом* ↔ *там*$_i$);
- от местоимения *это,* если у него есть именной антецедент (*...и они*$_i$ *сами, и вырытые ими окопы*$_j$ *— все это*$_{i+j}$ *было уже в тылу у немцев*).

2) кореферентные связи:

a. между неместоименными существительными, а также прилагательными и причастиями, которые выполняют функции существительных (*Тот черненький*$_i$*, интеллигентненький, что примечал его при регистрации как своего, прямо направился теперь к Монахову как к своему. <...> Он поискал глазами того*$_i$*, черненького, и не нашел*) или остаются после эллипсиса существительного-вершины в именной группе (*Одна пушка, брошенная всеми, завалилась набок ... <...> Около другой*$_i$ *суетился расчет. Они на руках скатывали ее*$_i$ *вниз, надеясь открыть огонь...*). Они могут быть предметными, событийными или обозначениями отрезков времени и иметь

любой референциальный статус, кроме предикативного, при котором именная группа не отсылает ни к какой сущности или классу сущностей (*Лева рос в так называемой академической среде и с детства мечтал стать ученым*; *Он ринулся в самую гущу толпы. И вышел победителем в борьбе за трап*). В частности, в нашем корпусе проводятся кореферентные связи между родовыми именными группами, отсылающими к одному и тому же классу (*Поэтому очень важно снижать металлоемкость полиграфической техники, тем самым добиваясь наибольшей экономии металла$_i$. А нельзя ли придумать такой способ печати, чтобы обходиться вообще без металла$_i$?*; *— Лес$_i$, оказывается, — с надеждой продолжил Монахов, — не просто много деревьев... И не каждое дерево, а лес$_i$ в целом существует, как единый организм...*);

b. между существительными и притяжательными прилагательными или некоторыми видами относительных прилагательных (*Люда ↔ Людин*; *Сибирь ↔ сибирский*);

c. между местоимениями 1-го и 2-го лица и другими упоминаниями (*Тогда Ищенко$_i$ закричал: — Ты что, проверяешь меня$_i$?*).

Разметка кореферентных связей в СинТагРусе проводится одним лингвистом в полуавтоматическом режиме: для каждого предложения сначала запускаются правила разрешения анафоры и кореферентности в системе ЭТАП-4, а если они устанавливают неправильные связи, разметчик исправляет их вручную (также вручную размечаются длинные связи, которые правила ЭТАП-4 не могут установить). После этого проводится автоматическая проверка ошибок разметки, а затем размеченный текст проверяется другим лингвистом.

На данный момент подкорпус СинТагРуса с кореферентной разметкой может использоваться в исследовательских целях, а в перспективе, по мере его увеличения, — и для обучения систем разрешения анафоры и кореферентности. О режиме доступа к данному подкорпусу см. https://ruscorpora.ru/page/corpora-datasets/.

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. Результаты и перспективы. М.: Индрик, 2005. С. 193–214.

Богуславский И. М., Иомдин Л. Л., Митюшин Л. Г., Сизов В. Г. Длина синтаксических связей в русском аннотированном корпусе // Международная конференция «Корпусная лингвистика — 2008», Санкт-Петербург, 2008. 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) // Труды Института русского языка им. В. В. Виноградова / Национальный корпус русского языка: 10 лет проекту. 2015. Вып. 6. С. 272–299.

Инструкция 2023: Синтаксическая разметка [Электронный ресурс]. URL: https://ruscorpora.ru/page/instruction-syntax

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. 2019. № 21. С. 14–41. DOI: 10.31912/pvrli-2019.21.1

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.

Савчук С. О. Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: Индрик, 2005. С. 62–88.

Тимошенко С. П., Иомдин Л. Л., Гладилин С. А., Иншакова Е. С. СинТагРус в составе НКРЯ: новые возможности // Труды международной конференции «Корпусная лингвистика-2021» (г. Санкт-Петербург, 01–03 июля 2021 г.). СПб.: Изд-во Санкт-Петербургского университета, 2021. С. 31–43.

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.

Boguslavsky I. SynTagRus — a Deeply Annotated Corpus of Russian // Les émotions dans le discours. Emotions in Discourse / ed. by P. Blumenthal, I. Novakova, and D. Siepmann. Peter Lang Edition, 2014. pp. 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. pp. 987–991.

Budnikov A., Toldova S., Zvereva D, Maximova D., Ionov M. Ru-Eval-2019: Evaluating anaphora and coreference resolution for Russian // Computational Linguistics and Intellectual Technologies. Supplementary Volume, 2019. pp. 2–13.

Chaga A. On a specific Russian construction with saturative verbs and negation // Annual International Conference "Dialogue" 2021, Student Session, Moscow, 2021.

Dobrovolskii V., Michurina M., Ivoylova A. RuCoCo: a new Russian corpus with coreference annotation // Computational Linguistics and Intellectual Technologies, 2022. pp. 141–149.

Endresen A. A., Zhukova V. A., Mordashova D. D., Rakhilina E. V., Lyashevskaya O. N. The Russian Constructicon: A New Linguistic Resource, Its Design and Key Characteristics // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2020). Issue 19 (26). pp. 241–255.

Fillmore Charles J., Kay Paul, O'Connor Mary Catherine. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone // Language. Vol. 64. No. 3 (Sep., 1988). P. 501–538.

Hajič J., Bejček E., Bémová A., Buráňová E., Hajičová E., Havelka J., Homola P., Kárník J., Kettnerová V., Klyueva N., Kolářová V., Kučová L., Lopatková M., Mikulová M., Mírovský J., Nedoluzhko A., Pajas P., Panevová J., Poláková L., Rysová M., Sgall P., Spoustová J., Straňák P., Synková P., Ševčíková M., Štěpánek J., Urešová Z., Vidová Hladká B., Zeman D., Zikánová Š., Žabokrtský Z. Prague Dependency Treebank 3.5 [Электронный ресурс]. URL: http://hdl.handle.net/11234/1-2621.

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. Microsyntactic Annotation of Corpora and its Use in Computational Linguistics Tasks // Jazykovedný časopis, ročník 86, číslo 2, 2017. pp. 169–178.

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

Toldova S., Roytberg A., Nedoluzhko A., Kurzukov M., Ladygina A., Vasilyeva M., Azerkovich I., Grishina Y., Sim G., Ivanova A., Gorshkov D. Evaluating Anaphora and Coreference Resolution for Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014). Issue 13 (20). pp. 681–695.

Uryupina O., Artstein R., Bristot A., Cavicchio F., Delogu F., Rodriguez K., Poesio M. Annotating a broad range of anaphoric phenomena, in a variety of genres: The ARRAU Corpus // Natural Language Engineering, 26 (1), 2020. pp. 95–128.

Igor M. Boguslavsky, Alexandra V. Chaga, Pavel V. Djachenko, Tatyana I. Frolova, Evgenia S. Inshakova, Leonid L. Iomdin, Alexander V. Lazurski, Leonid G. Mityushin, Andrey A. Movsesyan, Ivan P. Rygaev, Victor G. Sizov, Svetlana P. Timoshenko

A. A. Kharkevich Institute for Information Transmission Problems, RAS (Moscow)

igor.m.boguslavsky@gmail.com, chagachaga@gmail.com, pavel.v.djachenko@gmail.com, tfrolova@gmail.com, e.s.inshakova@gmail.com, iomdin@gmail.com, lazursky@mail.ru, lmityushin@gmail.com, derise@iitp.ru, irygaev@jent.ru, victor.sizov@gmail.com, nyrestein@gmail.com

THE CURRENT STATE OF THE SYNTAGRUS CORPUS

The paper presents a description of the main features and options of a diversely tagged corpus of Russian texts called SynTagRus. The corpus has been developed by the A. A. Kharkevich Institute for Information Transmission Problems, RAS, and is currently considered to be a subcorpus of RNC, where it is referred to as the "Syntactic Corpus".

Much attention is given to the linguistic principles underlying the different annotation types: morphological, syntactic, lexical semantic, lexical functional, elliptical, microsyntactic, coreferential, and temporal. Statistical data are given which characterize a variety of aspects of SynTagRus and its fragments. SynTagRus is a corpus with a 100-percent disambiguation at all levels of annotation. The paper outlines the obvious advantages of this approach but at the same time notes the difficulties associated with the need to always make definite decisions and choose single annotation options even in cases when the linguistic material undeniably allows for multiple linguistic description. Much attention is given to certain differences that exist between the SynTagRus and the main RNC subcorpora, such as distribution of words by parts of speech or specific morphological solutions that are accepted in SynTagRus in contrast to RNC (e.g. individual morphological categories, like verbal aspect and voice, certain cases of nouns etc.).

Keywords: SynTagRus, Syntactic Corpus, Morpho-syntactic Tagging, Lexical Tagging, Elliptical Tagging, Microsyntax, Coreference Tagging, Temporal Tagging

References

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: 2003–2005. Results and Prospects]. Moscow, Indrik Publ., 2005, pp. 193–214. (In Russ.)

Boguslavsky I. SynTagRus — a Deeply Annotated Corpus of Russian. *Les émotions dans le discours. Emotions in Discourse* / ed. by P. Blumenthal, I. Novakova, and D. 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, 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.)

Budnikov A., Toldova S., Zvereva D., Maximova D., Ionov M. Ru-Eval-2019: Evaluating anaphora and coreference resolution for Russian. *Computational Linguistics and Intellectual Technologies.* Supplementary Volume. Moscow, RSUH Publ., 2019, pp. 2–13.

Chaga A. On a specific Russian construction with saturative verbs and negation. *Annual International Conference "Dialogue" 2021*, Student Session, Moscow, 2021.

Dobrovolskii V., Michurina M., Ivoylova A. RuCoCo: a new Russian corpus with coreference annotation. *Computational Linguistics and Intellectual Technologies*. Moscow, RSUH Publ., 2022, pp. 141–149.

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.)

Endresen A. A., Zhukova V. A., Mordashova D. D., Rakhilina E. V., Lyashevskaya O. N. The Russian Constructicon: A New Linguistic Resource, Its Design and Key Characteristics. *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"* (2020). Issue 19 (26). Moscow, RSUH Publ., 2020, pp. 241–255.

Fillmore Charles J., Kay Paul, O'Connor Mary Catherine. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language*, vol. 64, no. 3 (Sep., 1988), pp. 501–538.

Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J., Poláková, L., Rysová, M., Sgall, P., Spoustová, J., Straňák, P., Synková, P., Ševčíková, M., Štěpánek, J., Urešová, Z., Vidová Hladká, B., Zeman, D., Zikánová, Š., Žabokrtský, Z. *Prague Dependency Treebank 3.5.* Available at: http://hdl.handle.net/11234/1-2621 (accessed 10.10.2023)

Inshakova E. S., Iomdin L. L., Mitjushin L. G., Sizov V. G., Frolova T. I., Cinman L. L. [SynTagRus today]. *Trudy Instituta russkogo jazyka im. V. V. Vinogradova* [Proc. of V. V. Vinogradov Russian Language Institute]. Vol. 21, Moscow, 2019, pp. 14–41. (In Russ.)

Instrukcija 2023: Sintaksicheskaja razmetka [Instruction 2023: Syntactic annotation]. Available at: https://ruscorpora.ru/page/instruction-syntax (accessed 10.10.2023)

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

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

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.)

Savchuk S. O. [Metatext markup in the National Corpus of the Russian Language: basic principles and main functions]. *Nacional'nyj korpus russkogo jazyka: 2003–2005. Rezul'taty i perspektivy*. [Russian National Corpus: 2003–2005. Results and prospects.]. Moscow, Indrik Publ., 2005, pp. 62–88. (In Russ.)

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.)

Timoshenko S. P., Iomdin L. L., Gladilin S. A., Inshakova E. S. [SynTagRus as a part of RNC: new perspectives]. *Trudy mezhdunarodnoi konferentsii "Korpusnaya lingvistika-2021"* [Proc. Int. Conf. "Corpus Linguistics-2021"]. Saint Petersburg, 2021, pp. 31–43. (In Russ.)

Toldova S., Roytberg A., Nedoluzhko A., Kurzukov M., Ladygina A., Vasilyeva M., Azerkovich I., Grishina Y., Sim G., Ivanova A., Gorshkov D. Evaluating Anaphora and Coreference Resolution for Russian. *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"* (2014). Issue 13 (20). Moscow, RSUH Publ., 2014, pp. 681–695.

Uryupina O., Artstein R., Bristot A., Cavicchio F., Delogu F., Rodriguez K., Poesio M. Annotating a broad range of anaphoric phenomena, in a variety of genres: The ARRAU Corpus. *Natural Language Engineering*, 26(1), 2020, pp. 95–128.