[1]П. В. Дяченко, [2]Л. Л. Иомдин, [3]А. В. Лазурский, [4]Л. Г. Митюшин,
[5]О. Ю. Подлесская, [6]В. Г. Сизов, [7]Т. И. Фролова, [8]Л. Л. Цинман
[1,2,3,4,5,6,7,8]Институт проблем передачи информации
им. А.А. Харкевича РАН
[1,2,3,4,5,6,7,8](Россия, Москва)
[1]pavelvd@iitp.ru, [2]iomdin@iitp.ru, [3]lazur@iitp.ru, [4]mit@iitp.ru,
[5]olga@iitp.ru, [6]sizov@iitp.ru, [7]tfrolova@cl.iitp.ru, [8]cinman@iitp.ru

СОВРЕМЕННОЕ СОСТОЯНИЕ ГЛУБОКО АННОТИРОВАННОГО КОРПУСА ТЕКСТОВ РУССКОГО ЯЗЫКА (СИНТАГРУС)

В статье излагаются основные особенности, принципы создания и параметры синтаксически аннотированного корпуса русских текстов «СинТагРус». Помимо синтаксической разметки каждого предложения в виде деревьев зависимостей, корпус содержит информацию об аргументах и значениях лексических функций слов, входящих в предложение, а также сведения о лексических значениях слов. Рассматривается подкорпус предложений, содержащих различные виды эллипсиса. Обсуждаются возможности применения корпуса для решения научных и практических задач.

Ключевые слова: СинТагРус, синтаксически размеченный корпус русских текстов, грамматика зависимости, лексические функции, разрешение неоднозначности, эллипсис.

Вводные замечания

Глубоко аннотированный корпус русских текстов (СинТагРус) в течение последнего десятилетия разрабатывается Лабораторией компьютерной лингвистики ИППИ РАН. Он основан на идеологии многоцелевого лингвистического процессора ЭТАП-3 (см., в част-

ности, Apresjan et al. 2003) и в первую очередь синтаксического анализатора (парсера) русского языка, используемого в различных приложениях лингвистического процессора, в том числе в системе русско-английского машинного перевода. СинТагРус – это составная, но полностью автономная часть Национального корпуса русского языка. Материалы корпуса (за исключением информации о дополнительных типах разметки, см. ниже раздел 2) доступны на сайте НКРЯ по адресу http://ruscorpora.ru/search-syntax.html.

Корпус СинТагРус содержит русские тексты, снабженные полной морфосинтаксической разметкой со снятой омонимией. Общий объем корпуса на конец 2014 года составляет около 63000 предложений (свыше 900 000 словоупотреблений) и постоянно растет. Большой объем корпуса и детальность содержащегося в нем материала делают его весьма ценным и уникальным ресурсом, который может быть успешно использован в широком классе исследовательских и практических задач компьютерной лингвистики (см. об этом, например, Иомдин, Иомдин 2013).

Тексты, включенные в СинТагРус, относятся к художественному, научно-популярному, публицистическому, биографическому и новостному жанрам. Выбор текстов этих жанров позволяет отразить в синтаксически аннотированном корпусе современное состояние русского литературного языка и в то же время ограничить появление в корпусе материала, синтаксическая, а частично и лексическая кодификация которого затруднительна (разговорная речь, поэзия, диалектные тексты, техническая документация). Как следствие, материал корпуса оказывается синтаксически достаточно однородным, что приводит к повышению качества синтаксической разметки и достижению большего единообразия принимаемых лингвистических решений.

В последние годы создатели корпуса ведут активную работу над обогащением синтаксически аннотированного корпуса лексико-функциональной разметкой. Лексические функции понимаются в соответствии с теорией лексических функций И. А. Мельчука и с учетом корректив, внесенных в нее Ю. Д. Апресяном. Около 13 000 предложений размечены лексическими функциями, осуществляется поиск по аргументам и значениям лексических функций, и в ближайшем будущем он также будет доступен для исследовательских целей на сайте НКРЯ.

Этапы развития корпуса и различные аспекты, связанные с его созданием, поддержкой и использованием, неоднократно публиковались в литературе (см., в частности, Boguslavsky et al., 2000; Апресян и др., 2005; Богуславский и др., 2008а,б; Шеманаева, Фролова, 2010; Boguslavsky 2014).

Настоящая статья отражает современное состояние СинТагРус'а. Статья содержит пять разделов. В первом разделе излагаются общие принципы построения и синтаксического аннотирования корпуса. Второй раздел посвящен дополнительным видам разметки корпуса: сведениям о лексических функциях и информации о наличии в предложении эллипсиса разных типов. В третьем разделе говорится о применении корпуса СинТагРус для регрессионного тестирования синтаксического парсера ЭТАП-3, а также для построения на его основе гибридного парсера русского текста. Четвертый раздел коротко характеризует исследования и разработки, выполненные в других коллективах, использующих СинТагРус. Наконец, пятый раздел посвящен ближайшим перспективам развития корпуса, в частности, автоматическому пополнению корпуса за счет части основного корпуса НКРЯ, а именно, подкорпуса со снятой омонимией.

1. Построение корпуса и принципы синтаксического аннотирования

Построение глубоко аннотированных корпусов текстов разных языков является весьма актуальной и быстро развивающейся областью современной компьютерной лингвистики. Это обусловлено, в частности, тем, что при создании систем автоматической обработки текстов все более широко применяются методы машинного обучения, и размеченные корпусы представляют собой именно те массивы данных, на которых такое обучение проводится. СинТагРус тут не является исключением; об использовании СинТагРус'а в машинном обучении различных компьютерно-лингвистических систем см. Богуславский и др. 2008а,б, Nivre et al. 2008, Казенников 2010, Казенников, Куракин 2011, Boguslavsky et al. 2011, Boguslavsky 2014.

Важной особенностью данного корпуса является представление синтаксических структур предложений в виде деревьев зависимостей. При этом различается около 70 типов синтаксических связей между словами, что обеспечивает более полное и лингвистически

содержательное представление синтаксической структуры, чем в других существующих корпусах с синтаксической разметкой.

В настоящее время существует не менее 70 корпусов для различных языков, в которых аннотация достигает синтаксического уровня, так называемых treebanks (в том числе не менее 15 для английского языка). СинТагРус занимает здесь особое место, являясь единственным в мире большим стопроцентно отредактированным корпусом текстов на русском языке с аннотацией на морфосинтаксическом уровне.

Корпус СинТагРус создается следующим образом. Вначале непрепарированный текст обрабатывается программой сегментации текста, которая автоматически разбивает его на отдельные предложения. После этого каждое предложение пропускается через парсер многофункционального лингвистического процессора ЭТАП-3, в результате чего вырабатывается его синтаксическая структура (в виде дерева зависимостей). Парсер выполняет также лексико-семантическую и лексико-функциональную (см. раздел 2.1) разметку предложения.

Чтобы лучше представить себе арсенал синтаксических решений, используемых в корпусе СинТагРус, рассмотрим примеры предложений, синтаксические структуры которых не относятся к числу самых простых.

На рис. 1 представлена синтаксическая структура предложения

(1) *Те детали, без которых прошлое не складывается в гармоничную картину, нарушает покой и расстраивает сон*

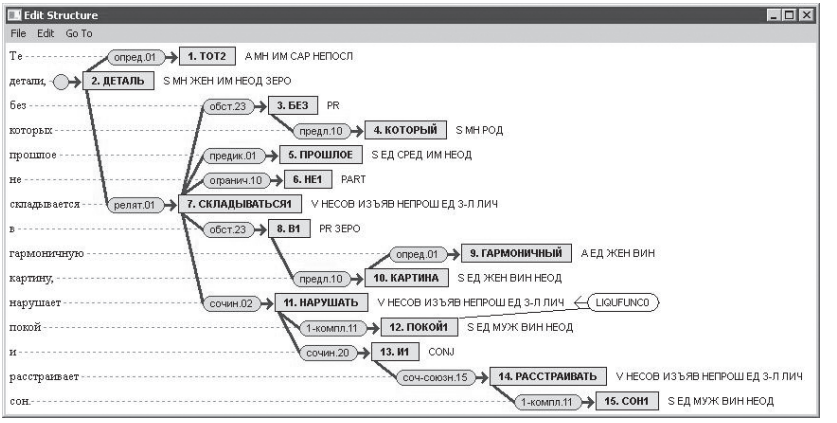

Рис. 1. Синтаксическая структура предложения (1).

Как видно из рисунка, в узлах 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). Как явствует из синтаксической структуры, предложение (1) — номинативное, его вершиной является существительное *детали*. Справа от него стоит придаточное определительное, вершина которого — личный глагол *складывается* — присоединяется к существительному по релятивной синтаксической связи. Этот глагол открывает сочинительную цепочку, куда входят еще два глагола — *нарушает* и *расстраивает*.

Лексико-семантическая разметка заключается в том, что многозначным словам приписывается одно из значений, имеющихся в комбинаторном словаре процессора ЭТАП-3. Так, в (1) обнаруживается лексическая многозначность, которая разрешается парсером (или лингвистом-экспертом при редакции), и мы видим результаты разрешения этой многозначности: в предложении фигурирует а) (местоименное) прилагательное ТОТ2, а не местоименное существительное ТОТ1 (ср. *Жалок тот, в ком совесть нечиста*), глагол СКЛАДЫВАТЬСЯ1 в значении ‘происходить, завершаться’ (ср. *Всё сложилось хорошо*) в отличие от глагола СКЛАДЫВАТЬСЯ2 в значении ‘состоять’ (как в *Роман складывается из нескольких частей*), союз И1, а не частица И2 (как в *Я и не знал, что он в городе*), существительное ПОКОЙ1 в значении ‘спокойное состояние’, а не существительное ПОКОЙ2 в значении ‘комната’ (ср. *царские покои*) и существительное СОН1 в значении ‘состояние с минимальным уровнем мозговой деятельности’ в отличие от существительного СОН2 в значении «сновидение».

На следующем рисунке представлена СинтС предложения

(2) *Что говорить: корпус действительно нуждается в капитальной реконструкции (это вовсе не означает его полный снос)*

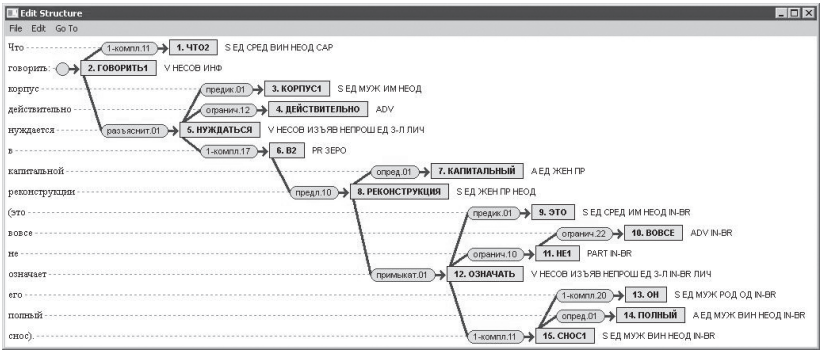

Рис. 2. Синтаксическая структура предложения (2).

Вершиной предложения (2) является глагол в инфинитиве *говорить*, лексическое значение которого ГОВОРИТЬ1 ‘сообщать’ (в отличие от ГОВОРИТЬ2 ‘разговаривать’, как в *говорить по-французски*). Этому глаголу подчиняется по 1-му комплетивному отношению местоименное существительное ЧТО2 в винительном падеже (ЧТО1 — это союз), тем самым констатируется, что *что* выступает в качестве прямого дополнения. Разумеется, это решение не отражает того факта, что *что говорить* — это своего рода синтаксическая фразема со значением ≈ ‘незачем говорить вследствие очевидности’, но на данном этапе синтаксически размеченный корпус такую информацию передать не может.

Отметим еще несколько деталей СинтС предложения (2): в нем фигурирует лексема КОРПУС2 ‘здание’ (а не КОРПУС2 ‘совокупность’ как в *корпус текстов*, КОРПУС3 ‘оболочка’ как *часы в золотом корпусе* и не КОРПУС3 ‘тело, туловище’, как в *подался всем корпусом*), лексема СНОС1 ‘разрушение’ (а не СНОС2 ‘смещение’, как в *снос корабля течением*); вводное предложение в скобках своей вершиной *означает* присоединяется к предшествующему существительному *реконструкции*.

(3) *«Теперь мы, пользуясь статистическими и компьютерными методами, будем изучать синтаксис и грамматическую структуру надписей из долины Инда», – говорит Раджеш Рао.*

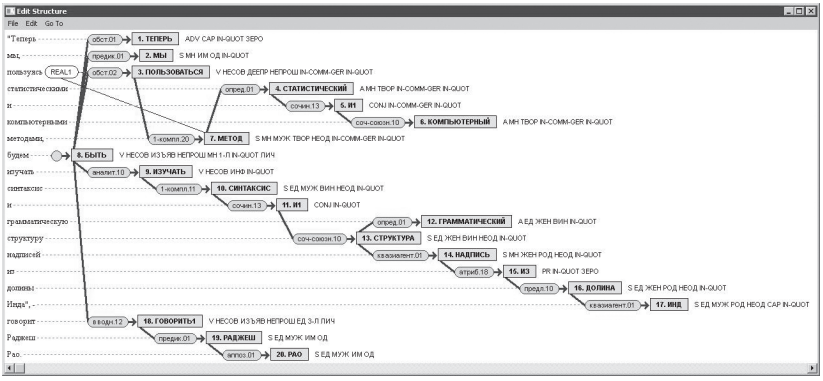

Рис. 3. Синтаксическая структура предложения (3).

Вершиной этого предложения является глагол *будем*, ему по вводному синтаксическому отношению подчиняется глагол со значением ГОВОРИТЬ1. Тут все дело в порядке слов: если бы глагол *говорит* предшествовал глаголу *быть*, то он был бы вершиной предложения и глагол быть подчинялся бы ему по 1-му комплетивному отношению (как в предложени*и Раджеш Рао говорит: теперь мы будем изучать синтаксис*). В данной фразе мы видим также примеры двух сочинительных конструкций — сочинение прилагательных статистическими и компьютерными (методами) и существительных, синтаксис и (грамматическую) структуру.

Фактически почти все затраты труда при разработке корпуса приходятся на последний этап работы, выполняемый специально подготовленными аннотаторами-лингвистами. Хотя лингвистический процессор ЭТАП-3 насыщен весьма богатой и разнообразной лингвистической информацией, он не способен строить морфосинтаксические структуры со стопроцентной надежностью. Аннотаторы проверяют все элементы структур, созданных парсером процессора, и вносят необходимые коррективы с помощью специального программного комплекса «Редактор структур» (основная составная часть программной среды «Рабочее место аннотатора»).

Использование процессора ЭТАП-3 для разработки корпуса создает обратную связь, важную для самого процессора, поскольку при выполнении этой работы база русских грамматических правил и русский словарь процессора постоянно пополняются и совершен-

ствуются. Отметим, что процессор рассчитан на обработку русских и английских текстов без каких-либо тематических и лексико-грамматических ограничений. Русский морфологический словарь в настоящее время содержит около 130 000 входов, а комбинаторный словарь — более 100 000 входов.

2. Дополнительные виды разметки корпуса

2.1. Лексико-функциональная разметка.

В ходе лексико-функциональной разметки обнаруживаются и отмечаются словосочетания, допускающие интерпретацию в терминах лексических функций; при этом также используется информация, содержащаяся в комбинаторном словаре. Затем структуры, построенные системой ЭТАП-3, проверяются экспертами-лингвистами.

Главный тезис теории лексических функций звучит так: в языке можно выделить несколько десятков значений высокого уровня абстракции (‘высокая степень’, ‘начало’, ‘прекращение’, ‘каузация’, ‘ликвидация’, ‘манифестация’ и т.п.), каждое из которых выражается большим классом слов. При этом выбор конкретного слова W для выражения данного значения ‘S’ целиком зависит от того слова X (аргумента ЛФ), с которым оно сочетается в тексте. Он идиоматичен, т.е. семантически не полностью мотивирован.

Маркировка ЛФ непосредственно в текстах и удобные средства поиска обеспечивают возможность прямого наблюдения контекстов, в которых реализуются ЛФ. Эти данные также важны для компьютерных систем переработки текстов на естественных языках. Системы, использующие ЛФ как инструмент анализа и перевода текстов, обладают преимуществом по сравнению с системами, которые их не используют. Кодируя универсальные смыслы, ЛФ становятся своеобразным языком-посредником между разными языками и позволяют описывать значительную часть лексики каждого рабочего языка независимо от других языков. Между тем, в большинстве систем машинного перевода правила перевода носят принципиально двуязычный характер и требуют постоянной корректировки уже имеющихся словарных данных при включении в систему новых языков.

В СинТагРус’е лексико-функциональные связи отражаются отдельно от синтаксической структуры с помощью особых элементов

XML. В визуализаторе же для удобства пользователя имена ЛФ записываются прописными латинскими буквами в овалах, от которых протянуты стрелки к аргументу и значению функции.

За примером обратимся снова к предложению (1). Нетрудно убедиться в том, что в этом предложении имеется несвободное словосочетание *нарушает покой*, которое описывается в терминах лексических функций. Из рис. 1 видно, что это словосочетание оформляется с помощью ЛФ LIQUFUNC0 от аргумента ПОКОЙ1 со значением НАРУШАТЬ. Аналогичным образом, из рис. 3 видно, что в предложении (3) встречается несвободное словосочетание *пользуясь методами*, которое оформляется с помощью ЛФ REAL1 от аргумента МЕТОД со значением ПОЛЬЗОВАТЬСЯ.

В номинативном предложении

(4) … *Который ставит под серьезную угрозу принцип равенства перед законом людей с разной сексуальной ориентацией*

отмечены два лексико-функциональных сочетания: *серьезную угрозу*, где СЕРЬЕЗНЫЙ — значение ЛФ MAGN от аргумента УГРОЗА2, и *ставить под угрозу*, где СТАВИТЬ — значение ЛФ INCEPLABOR1–2 от того же аргумента УГРОЗА2, а предлог ПОД1 — обязательный контекст этой лексико-функциональной зависимости.

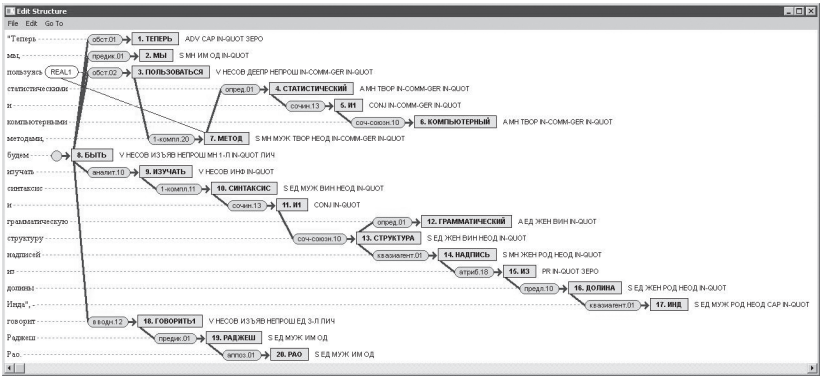

Рис. 4. Разметка предложения (4) несколькими лексическими функциями.

2.2. Обработка эллиптических конструкций

Одной из наиболее трудных задач, возникающих при автоматическом синтаксическом анализе, является обработка эллиптических конструкций.

Предложения с эллипсисом составляют 2–3 процента от общего числа предложений корпуса СинТагРус. Они представлены в корпусе деревьями зависимостей, в которых эллиптированным словам соответствуют особые узлы дерева, имеющие пустой текстовый элемент («фантом»). Специально для корпуса, для лучшего восприятия пользователями, был предложен ряд решений относительно формы представления эллиптических конструкций.

Упомянем два таких решения.

1) Восстановление некоторых типов синтаксического эллипсиса

В случаях, когда опущенные слова физически присутствуют в другой части предложения (например, в сочинительной конструкции с сочинительным сокращением), эти слова восстанавливаются. Например, в предложении

(5) *И приходится пересаживать детям почки, сердце от взрослых и только в подростковом возрасте, когда организм готов принять большой орган*

между словами *и* и *только* вставляется узел «ПЕРЕСАЖИВАТЬ» с соответствующим набором характеристик и пустым текстовым элементом. От таких «фантомных» слов проводятся все необходимые связи (см. рис. 5):

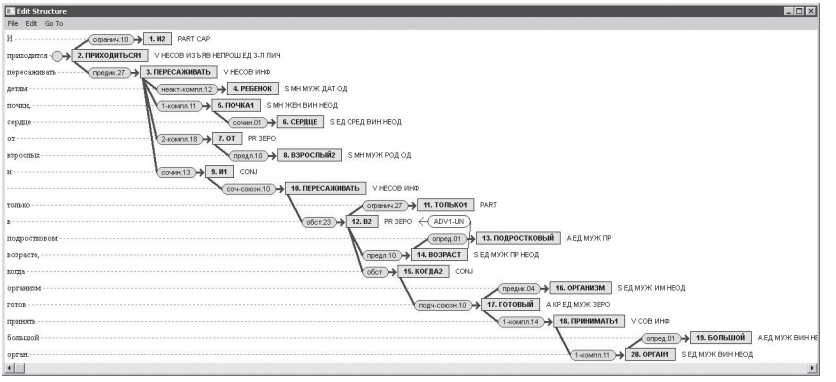

Рис. 5. Синтаксическая структура предложения (5) с восстановленным эллипсисом

При отсутствии восстановленного узла древесная структура предложения выглядела бы неестественно, а соответствующие синтаксические отношения невозможно было бы разумно интерпретировать.

Леммы восстановленных «фантомных» узлов совпадают с теми, которые уже встречались в предложении, а отдельные морфологические характеристики, например, род у глаголов прошедшего времени, могут меняться.

(6) *Отсюда он летит в Сочи, и сразу обратно в Уфу, на чемпионат Башкирии, а 5 апреля стартует чемпионат России в Ханты-Мансийске*

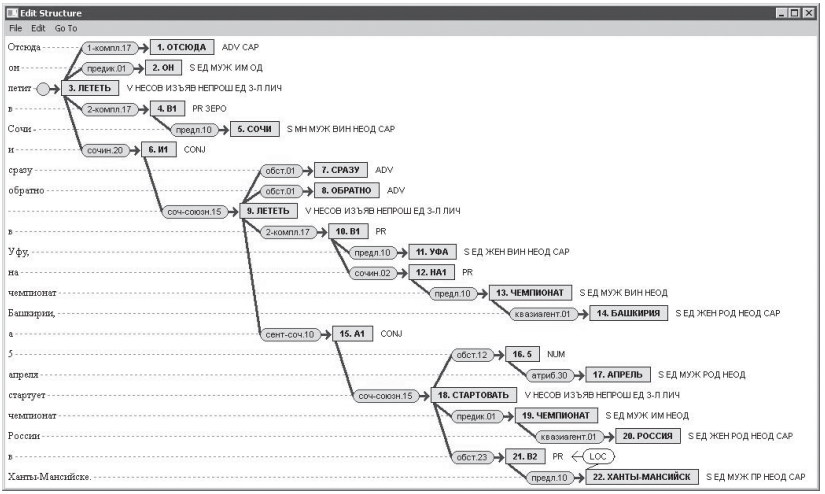

Рис. 6. Синтаксическая структура предложения (6) с восстановленным эллипсисом

В СинТагРус'е встречаются и достаточно курьезные примеры, понятные человеку из контекста, но не восстанавливаемые автоматически. Для того, чтобы получить приличную структуру второй фразы в тексте

(7) *Например, пшеничная мука первого сорта с 12 июня прошлого года до 4 марта нынешнего выросла в цене на 34 процента. Высшего — на 44%, ржанообдирочная мука — на 73%.*

пришлось, опираясь на данные предыдущей фразы из этого текста, дописать 8 «фантомных» узлов, среди них не только глаголы, но и имена существительные и предлоги:

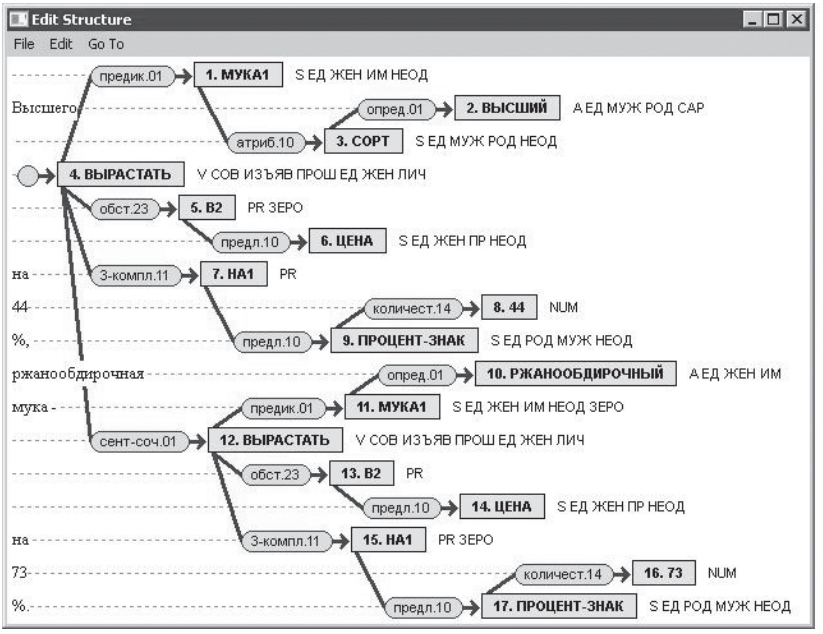

Рис. 7. Множественные восстановленные узлы в высокоэллиптичном предложении

Приведём для ясности восстановленную фразу целиком; здесь фантомные узлы написаны светлым шрифтом, а реально фигурирующие слова выделены жирным шрифтом:

(7′) *Мука* ***Высшего*** *сорта выросла в цене —* ***на 44%, ржанообдирочная мука*** *выросла в цене —* ***на 73%****.*

2) Использование специальных фиктивных слов для частичной нормализации высокоэллиптичных предложений

Этот прием используется в тех случаях, когда в предложении «опущен» глагол некоторой размытой семантики, так называемый неопределенный глагол, как, например, в предложении

(8) *Ждешь беду отсюда, ан из-за угла тебя мешком*

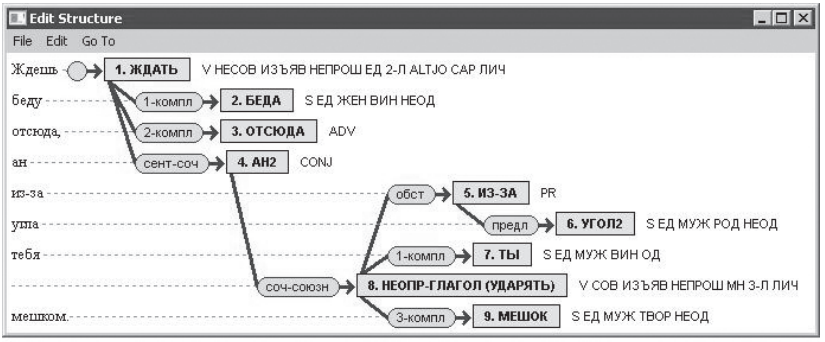

Рис. 8. Синтаксическая структура (8) с «фантомным» глаголом размытой семантики «ударять»

Приведем еще два примера фраз с эллипсисом, фигурирующих в корпусе:

(9) — *Как сами, так делают что хотят, а порядочному человеку — выговор*

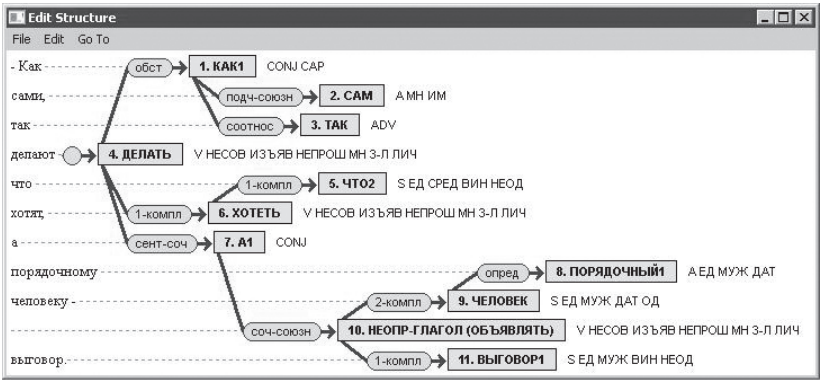

Рис. 9. Синтаксическая структура (9) с «фантомным» глаголом размытой семантики «объявлять»

(10) *Но это, когда разведчик уже лежал на снегу*

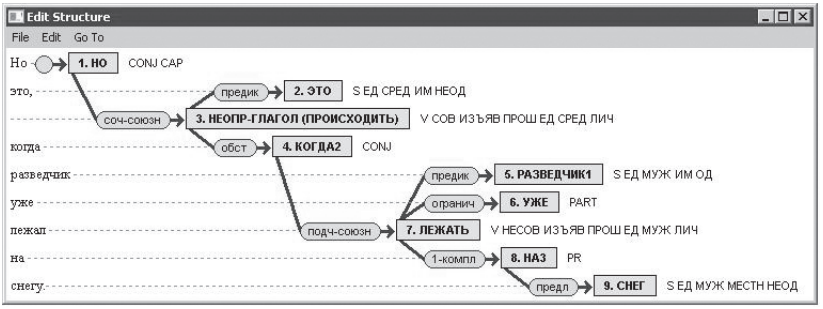

Рис. 10. Синтаксическая структура с «фантомным» глаголом размытой семантики «происходить»

Во всех таких случаях в предложение добавляется глагольная вершина, ей приписываются наиболее подходящие и естественные грамматические характеристики, а в качестве леммы пишется НЕОПР-ГЛАГОЛ (неопределенный глагол) и затем в скобках глагол, который является уместной и естественной гипотезой. Так, в предложении (8) после слова *тебя* добавляется узел с леммой НЕОПР-ГЛАГОЛ (УДАРЯТЬ), в предложении (9) после слова *человеку* — НЕОПР-ГЛАГОЛ (ОБЪЯВЛЯТЬ), а в последнем предложении после *это* — узел с леммой НЕОПР-ГЛАГОЛ (ПРОИСХОДИТЬ).

3. Использование корпуса СинТагРус в лингвистическом процессоре ЭТАП-3.

3.1. Регрессионное тестирование парсера ЭТАП-3

Как было показано в разделе 1, важнейшим этапом построения корпуса является создание полных древесных синтаксических структур каждого предложения с помощью автоматического синтаксического анализатора (парсера) системы ЭТАП-3, которые затем проверяются и при необходимости корректируются лингвистами-экспертами. При этом оказывается, что часть синтаксических структур не требует коррекции и вносится в корпус по результатам проверки безо всяких изменений. По мере развития корпуса и модернизации самого парсера доля таких предложений в корпусе постоянно увеличивается и сейчас составляет около 35% (приблизительно 20 000 предложений). Совокупность таких предложений

удобно использовать в качестве своего рода золотого стандарта – эталона, относительно которого удобно тестировать текущее состояние парсера системы ЭТАП-3 и ресурсов, на которые он опирается (словари и различные совокупности правил).

Вкратце процедура регрессионного тестирования парсера с помощью корпуса СинТагРус (впервые предложенная в работе Богуславский и др., 2008) выглядит следующим образом. Время от времени весь корпус пропускается через парсер системы ЭТАП-3. Результаты работы парсера автоматически анализируются, и выделяется эталонная совокупность предложений, структуры которых полностью совпадают со структурами отредактированного экспертами корпуса. Эту совокупность предложений мы называем текущим регрессионным тестом. Далее, после каких-либо изменений, внесенных в словари или правила, которые разработчики считают чувствительными (это может быть добавление новых слов или новых значений слов, редактирование существующих или создание новых правил), текущий регрессионный тест пропускается через парсер системы ЭТАП-3 в его актуальном состоянии. Автоматическое сравнение результатов работы предшествующего состояния парсера и его актуального состояния на текущем регрессионнном тесте позволяет увидеть все последствия сделанных изменений. Некоторые из этих изменений являются нежелательными и оказываются следствием ошибочных манипуляций с ресурсами парсера, которые приходится пересматривать; заметим, что иногда такие изменения носят просто-таки разрушительный характер, и своевременное их устранение жизненно необходимо. В то же время другая часть изменений, напротив, является желательной: в результате учета этих изменений совершенствуется как корпус, так и парсер системы, которые оказываются благотворно влияющими друг на друга. Подчеркнем, что объем текущего регрессионного теста вполне достаточен, чтобы он уверенно “откликался” даже на небольшие изменения ресурсов корпуса, т.е. словарей и правил, а также на изменения в алгоритмическом и программном обеспечении системы. Наконец, регрессионное тестирование позволяет оптимизировать выбор опций и настроек парсера, которые оказываются наиболее удобными для создания размеченного корпуса.

3.2. Создание гибридного парсера на основе парсера ЭТАП-3 и СинТагРус'а.

В последние несколько лет активно развивается подход, связанный с использованием статистических данных корпуса СинТагРус непосредственно в парсере лингвистического процессора ЭТАП-3. На основании данных корпуса вырабатываются оценки вероятностей альтернативных вариантов для элементов морфосинтаксической структуры, а именно оценки вероятностей различных лексико-морфологических интерпретаций одного и того же слова предложения и оценки вероятностей различных синтаксических связей, входящих в один и тот же узел дерева зависимостей. На основе этих данных парсер ищет дерево зависимостей, удовлетворяющее всем требованиям грамматических правил системы и имеющее максимальную результирующую оценку вероятности. Парсер ЭТАП-3, поддерживаемый статистикой корпуса, разумно считать своего рода гибридной системой, использующей комбинацию правилового подхода, на котором базируется исходный парсер ЭТАП-3, и статистическими данными.

Как показали масштабные эксперименты, гибридный подход ощутимо повышает качество синтаксических структур, получаемых парсером (см. Boguslavsky et al. 2011, 2014). Однако для дальнейшего совершенствования гибридного парсера требуется дополнительная исследовательская и экспериментальная работа.

4. Исследовательские задачи, решаемые с участием корпуса СинТагРус

По специальному лицензионному соглашению около 50 исследовательских групп и институтов получили в свободное пользование материалы корпуса СинТагРус.

Следует прежде всего отметить, что СинТагРус был использован для создания успешного работающего статистического Malt-парсера для русского языка (Nivre et al. 2008, Bohnet et al. 2013).

Сотрудники Исследовательского центра Samsung использовали СинТагРус для обучения и тестирования алгоритма по разрешению омонимии при помощи статистических методов. Статистические методы с успехом применялись, в частности, к извлечению темпоральных выражений и анализу именных групп, при этом подкор-

пус для машинного обучения был составлен на базе СинТагРус (Kudinov et al. 2014, Muzychka et al. 2014).

В компании Google СинТагРус и другие аналогичные корпусы текстов, размеченные синтаксическими зависимостями (для разных языков) использовали в целях обучения и тестирования многоязычного статистического парсера. Сообщается о том, что благодаря размеченным корпусам относительную ошибку анализа удавалось в ряде задач уменьшить на 13%, а в случае задачи распознавания именованных сущностей – даже на 26% (Täckström et al. 2012).

Участники норвежской исследовательской группы CLEAR, в задачи которой входит изучение русского языка, Х. Экхофф и А. Бердичевский (Норвежский Арктический университет, г. Тромсё) в рамках проекта Birds&Beasts занимаются разметкой и исследованиями на материале синтаксического корпуса TOROT (старославянский, древнерусский, старорусский): https://nestor.uit.no/ и сравнением данных, полученных по этому корпусу, с современным состоянием языка, представленном в СинТагРус'е. В частности, составляются синтаксические профили русских глаголов.

Количественное исследование синтаксического поведения некоторых классов русских слов было проведено на материале корпуса СинТагРус китайскими исследователями Ван Юном, Лю Хайтао (2013).

СинТагРус используется для целей классификации различных лингвистических объектов. Например, в работе Баранова 2013 корпус был использован для машинного обучения системы, определяющей уровень сложности русских предложений для целей обучения русскому языку как иностранному.

Наконец, синтаксически размеченный корпус русского языка может быть использован для составления лингвистических задач, направленных на поиск специальных конструкций (Шеманаева 2009).

5. Новейшее развитие корпуса СинТагРус и ближайшие перспективы

В последнее время разработчиками корпуса был проведен масштабный эксперимент по переводу фрагмента основной части Национального корпуса русского языка со снятой омонимией в формат СинТагРус'а. Цель эксперимента была двоякой: с одной стороны, планировалось проверить, насколько трудоёмкой и автоматизируе-

мой окажется работа по переводу морфологической разметки НКРЯ в морфологический формат СинТагРус'а, учитывая заметные расхождения морфологических трактовок обоих корпусов, а, с другой стороны, выяснить, насколько и как именно дополнительная информация о разрешенной неоднозначности лексических и морфологических элементов текста, присутствующая в НКРЯ, повлияет на качество синтаксического анализа, осуществляемого парсером ЭТАП-3. Содержание этой работы было подробно освещено в недавней статье Дяченко, Подлесская, Сизов, 2014; здесь излагаются ее основные результаты.

Следует отметить, что работа по переводу материалов размеченного корпуса из одного формата в другой предпринимались неоднократно, в том числе и с участием СинТагРус'а; ср., в частности, эксперименты по переводу СинТагРус'а в формализм HPSG (Avgustinova T., Zhang Y, 2010; Avgustinova 2011), а также эксперименты по переводу СинТагРус'а в формат Пражского корпуса зависимостей (PDT) для чешского языка (Mareček, Kljueva 2009), однако впервые материалы другого ресурса переводились в формат СинТагРус'а.

Мы исходили из предположения, что использование дополнительной информации такого рода в процессе автоматического построения синтаксической структуры предложения позволяет повысить качество получаемой структуры, тем более что у нас был успешный опыт аналогичного эксперимента с английским языком (Диконов, Дяченко 2010). В качестве источника информации использовался фрагмент НКРЯ, содержащий сведения о морфологическом разборе входящих в него слов со снятой вручную омонимией (будем для краткости называть эту часть НКРЯ Подкорпусом). Эксперимент проводился на фрагменте Подкорпуса, состоящем из текстов, написанных после 1950 г. (примерно 4,48 млн. словоупотреблений).

Между морфологическими разметками НКРЯ и СинТагРус'а существуют различия, которые не сводятся к простой замене имен меток, в то же время их можно преодолеть. В состав разметки подкорпуса входят морфологические характеристики и леммы. Мы рассчитывали, что учет этой информации поможет парсеру ЭТАП-3 разрешить морфологическую омонимию (в идеале — полностью), и тем самым существенно уменьшить число рассматриваемых аль-

тернативных вариантов разбора. Из оценки исходных данных следует, что около 83% неоднозначных разборов различаются морфологическими характеристиками или леммами и тем самым могут быть разрешены с использованием информации, взятой из Подкорпуса.

Оценить эффект данных из Подкорпуса на объем неоднозначностей, оставшихся после процедуры их разрешения, можно из следующей таблицы.

Словоупотребления, для которых неоднозначность снята полностью	1805555
Словоупотребления, для которых неоднозначность снята частично	21041
Словоупотребления, для которых ни один разбор парсера ЭТАП-3 не совпал с данными из Подкорпуса	171621
Всего словоупотреблений с неоднозначными результатами анализа	2507953

Табл. 1. Результаты разрешения неоднозначности.

Как видим, использование данных Подкорпуса позволяет уменьшить число неоднозначных морфологических разборов на 71%. Уменьшение числа неоднозначных разборов ожидаемо приводит к сокращению времени работы системы ЭТАП-3 с использованием данных Подкорпуса за счет сокращения числа рассматриваемых альтернативных гипотез, а именно: работа на фрагменте подкорпуса без учета его данных составила 37 ч. 15 м., а с учетом данных — 13 ч. 48 м.

Еще более показательным является сравнение качества синтаксического анализа некоторого фрагмента текста, использующего данные Подкорпуса, со стандартным СА.

Режим эксперимента	С данными Подкорпуса	Без данных Подкорпуса
Построено структур	255	
Сравнивалось слов	2740	2743
С верными частями речи и морфологическими характеристиками:	2632 (96%)	2528 (92%)
Верных лемм:	2570 (94%)	2519 (92%)

Режим эксперимента	С данными Подкорпуса	Без данных Подкорпуса
Верных структур (без учёта имён связей):	151 (60%)	148 (58%)
Верных структур (с учётом имён связей):	129 (50%)	123 (48%)

Табл. 2. Параметры качества синтаксического анализа

Приведенные в таблице 2 результаты показывают, что использование Подкорпуса при синтаксическом анализе приводит к существенному сокращению в разметке числа слов с неправильными морфологическими характеристиками, а также положительно влияет на качество построенных синтаксических структур. Правда, это изменение не столь значительно, как мы предполагали на ранних стадиях эксперимента. Возможно, на результат оказывают влияние еще не обнаруженные расхождения в разметке, которые, встречаясь достаточно редко, тем не менее ухудшают результат работы экспериментального синтаксического анализатора.

Достаточно высокое качество морфологического компонента и созданного на его основе корпуса позволит использовать результаты для обучения статистических программ, использующих морфологическую разметку — например, создать программу статистического морфологического анализа текстов в формате СинТагРус'а. В то же время недостаточно высокое качество построенных синтаксических структур, близкое к качеству результата работы ЭТАП-3 в обычном режиме, не позволяет применить полученный корпус в имеющемся виде для обучения статистического синтаксического анализатора, а также не дает существенного уменьшения трудозатрат при ручной разметке экспертами-лингвистами. Поскольку в ходе экспериментов были выявлены определенные резервы по усовершенствованию алгоритмов, то дальнейшие исследования, как мы надеемся, изменят ситуацию к лучшему.

Заметим, наконец, что в ближайшие планы разработчиков корпуса входит внесение разметки о различных микросинтаксических конструкциях русского языка и синтаксических фраземах. В результате СинТагРус окажется полезным и для исследований в области малого синтаксиса и фразеологии.

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.

Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л., Санников В.З. Теоретические проблемы русского синтаксиса: Взаимодействие грамматики и словаря. Отв. ред. Ю.Д. Апресян. М.: Языки славянских культур, 2010.

Баранова Ю.Н. Автоматическая классификация типов предложений (на базе СинТагРус). Данные междисциплинарного научного семинара (МНС) «Методы интеллектуального анализа естественного языка», Нижний Новгород, 2013.

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 научная конференция «Корпусная лингвистика», СПб., 2008а.

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

Диконов В.Г., Дяченко П.В. Эксперимент по построению синтаксической структуры английских предложений с использованием заранее известных фрагментарных данных // Информационные технологии и системы (ИТиС'10). Сборник трудов 33-ой Конференции молодых ученых и специалистов ИППИ РАН. Геленджик, 18-26 сентября 2010 г. М.: ИППИ, 2010. С. 310-319. ISBN 978-5-901158-12-8.

Дяченко П.В., Подлесская О.Ю., Сизов В.Г. НКРЯ: основной корпус и СинТагРус, синтаксический анализ текстов со снятой морфологической омонимией. // Информационные технологии и системы (ИТиС'2014). Сборник трудов 38-ой Конференции молодых ученых и специалистов ИППИ РАН. Нижний Новгород, Россия, 2014.

Иомдин Л.Л., Иомдин Б.Л. Отрицание и валентности в русском языке (по корпусным данным). // Международная научная конференция «Корпусная лингвистика». СПб.: С.-Петербургский гос. университет, 2013. С. 281–291.

Казенников А.О. Сравнительный анализ статистических алгоритмов синтаксического анализа на основе деревьев зависимостей. // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (2010) Выпуск 9 (16). С. 157–162.

Казенников А.О., Куракин Д.В. Сравнительный анализ алгоритмов синтаксического разбора в рамках задачи анализа новостных потоков. // Труды XVIII Всероссийской научно-методической конференции Телематика '2011, СПб, 2011. С. 235–236.

Шеманаева О.Ю. Глубоко аннотированный корпус русских текстов как обучающий электронный ресурс // Информационные технологии и системы (ИТиС'09). Сборник трудов 32-ой Конференции молодых ученых и специалистов ИППИ РАН. Бекасово, 15–18 декабря 2009 г. М., 2009. С. 205–209.

Шеманаева О.Ю., Фролова Т.И. Лексико-функциональная разметка текстов в СинТагРус // Информационные технологии и системы (ИТиС'10). Сборник трудов 33-й конференции молодых ученых и специалистов ИППИ РАН. Геленджик, 18-26 сентября 2010 г. М.: ИППИ, 2010. С. 320–324.

Ван Юн, Лю Хайтао. Квантитативное исследование имени существительного в русском языке по его синтаксическим признакам. // Вестник Московского университета. Сер. 9. Филология, 2013 (5). С. 35–51.

Avgustinova T. Parallel Construction of Slavic Grammatical Resources. Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог'2011». М.: Изд-во РГГУ, 2011. Вып. 10(17). С. 41–50.

Avgustinova T., Zhang Y. Conversion of a Russian Dependency Treebank into HPSG Derivations. Proceedings of the 9 th International Workshop on Treebanks and Linguistic Theories (TLT'9), 2010.

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, Ecole Normale Superieure, Paris, June 16-18 2003, pp. 279–288.

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. // Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000). San Francisco: Kaufmann, pp. 987–991.

Boguslavsky I., Iomdin L., Timoshenko S., Frolova T. Development of the Russian Tagged Corpus with Lexical Functional Annotation // Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop. Proceedings. Bratislava, Slovakia. April 15-16, 2009. Bratislava, 2009, p. 83–90.

Boguslavsky I., Iomdin L., Tsinman L., Petrochenkov V. Rule-Based Dependency Parser Refined by Empirical and Corpus Statistics // Proceedings of the International Conference on Dependency Linguistics (Depling'2011). Barcelona, September 5-7, 2011, p. 318–327.

Boguslavsky I., Iomdin L., Petrochenkov V., Sizov V., Tsinman L. A Case of Hybrid Parsing: Rules Refined by Empirical and Corpus Statistics // In: Computational Dependency Theory. Vol. 258 of Frontiers in Artificial Intelligence and Applications. Eds. Kim Gerdes, Eva Hajicova, Leo Wanner, 2014. pp. 226-240. IOS Press, ISBN 978-1-61499-351-3 DOI 10.3233/978-1-61499-352-0-226.

Bohnet B., Nivre J., Farkas R., Ginter F., Hajic J. Joint Morphological and Syntactic Analysis for Richly Inflected Languages // Transactions of the Association for Computational Linguistics, 2013. Vol. 1. pp. 415–428.

Iomdin L. Automatic Text Processing and Deeply Annotated Text Corpora of Russian: Interaction and Mutual Impact. // Jazykovedné štúdie XXXI. Bratislava: Jazykovedný ústav Ľ.Štúra Slovenskej akadémie vied, 2014. pp. 136–146.

Kudinov M., Romanenko A., Piontkovskaya I. Conditional Random Field in Segmentation and Noun Phrase Inclination on Tasks for Russian // Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог'2014». М.: 2014. С. 315–324.

Mareček D., Kljueva N. Converting Russian Treebank SynTagRus into Praguian PDT Style // Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages 2009, Bulgaria, pp. 26–31.

Muzychka S., Romanenko A., Piontkovskaya I. Conditional Random Field for Morphological Disambiguation in Russian // Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог'2014». М.: РГГУ, 2014. С. 474–483.

Nivre J., Boguslavsky I., Iomdin L. Parsing the SynTagRus Treebank of Russian. // In Proceedings of the 22nd International Conference on Computational Linguistics (COLING08), Scott D., Uszkoreit H. (eds.), 2008, pp. 641–648.

Podlesskaja O., Frolova T. Tagging lexical functions in Russian texts of SynTagRus // Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог'2011». М.: Изд-во РГГУ, 2011. Вып. 10(17). С. 207–218.

Täckström O., McDonald R. and Uszkoreit H. Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure // North American Association for Computational Linguistics (NAACL), 2012.

[1]P.V. Dyachenko, [2]L.L. Iomdin, [3]A.V. Lazursky, [4]L.G. Mityushin, [5]O.Yu. Podlesskaya, [6]V.G. Sizov, [7]T.I. Frolova, [8]L.L. Tsinman
[1,2,3,4,5,6,7,8]A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences
[1,2,3,4,5,6,7,8](Russia, Moscow)
[1]pavelvd@iitp.ru, [2]iomdin@iitp.ru, [3]lazur@iitp.ru, [4]mit@iitp.ru, [5]olga@iitp.ru, [6]sizov@iitp.ru, [7]tfrolova@cl.iitp.ru, [8]cinman@iitp.ru

A DEEPLY ANNOTATED CORPUS OF RUSSIAN TEXTS (SYNTAGRUS): CONTEMPORARY STATE OF AFFAIRS

The paper discusses the main features, principles of creation and parameters of the syntactically tagged corpus of Russian texts, SynTagRus. In addition to syntactic annotation of all sentences with dependency trees, the corpus contains information on arguments and values of lexical functions for words occurring in the sentence, as well as the data on lexical meanings of words. The article considers the subcorpus of sentences containing diverse types of ellipsis and discusses possible uses of the corpus for research tasks and applications.

Key words: SynTagRus, syntactically tagged corpus of Russian texts, dependency grammar, lexical functions, ambiguity resolution, ellipsis

References

Apresjan Ju.D., Boguslavskij 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* [Russian National Corpus: 2003–2005. Results and Prospects]. Moscow, Indrik Publ., 2005. (In Russ.)

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, Ecole Normale Superieure, Paris, June 16-18 2003, pp. 279–288.

Apresjan Ju.D., Boguslavskij I.M., Iomdin L.L., Sannikov V.Z. *Teoreticheskie problemy russkogo sintaksisa: Vzaimodeistvie grammatiki i slovarya* [Theoretical issues of Russian syntax: Interaction of the grammar and the dictionary]. Ju.D. Apresjan (Ed.). Moscow, Yazyki slavyanskikh kul'tur Publ., 2010.

Avgustinova T. Parallel Construction of Slavic Grammatical Resources. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2011)*. Issue 10 (17), Moscow: RSUH Publishers, pp. 41-50.

Avgustinova T., Zhang Y. Conversion of a Russian Dependency Treebank into HPSG Derivations. *Proceedings of the 9th International Workshop on Treebanks and Linguistic Theories (TLT'9)*, 2010.

Baranova Ju.N. [An automatic classification of sentence types (based on SynTagRus)]. *Dannye mezhdistsiplinarnogo nauchnogo seminara (MNS) "Metody intellektual'nogo analiza estestvennogo yazyka"*, Nizhnij Novgorod, 2013. (In Russ.)

Boguslavsky I. SynTagRus – a Deeply Annotated Corpus of Russian. *Les èmotions dans le discours. Emotions in Discourse. Peter Blumenthal,* Iva Novakova, Dirk Siepmann (Eds.). 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. *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. San Francisco, Kaufmann, 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 nauchnaya konferentsiya "Korpusnaya lingvistika"*, St.-Petersburg, 2008b. (In Russ.)

Boguslavsky I., Iomdin L., Timoshenko S., Frolova T. Development of the Russian Tagged Corpus with Lexical Functional Annotation. *Metalanguage and Encoding Scheme Design for Digital Lexicography.* MONDILEX Third Open Workshop. Proceedings. Bratislava, Slovakia. April 15-16, 2009. Bratislava, 2009, pp. 83–90.

Boguslavsky I., Iomdin L., Petrochenkov V., Sizov V., Tsinman L. A Case of Hybrid Parsing: Rules Refined by Empirical and Corpus Statistics. *Computational Dependency Theory*. Vol. 258 of Frontiers in Artificial Intelligence and Applications. Eds. Kim Gerdes, Eva Hajicova, Leo Wanner, 2014. pp. 226-240. IOS Press, ISBN 978-1-61499-351-3 DOI 10.3233/978-1-61499-352-0-226.

Boguslavsky I., Iomdin L., Tsinman L., Petrochenkov V. Rule-Based Dependency Parser Refined by Empirical and Corpus Statistics. *Proceedings of the International Conference on Dependency Linguistics* (Depling'2011). Barcelona, September 5–7, 2011, pp. 318–327.

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 nauchnaya konferentsiya "Korpusnaya lingvistika"*, St.-Petersburg, 2008a. (In Russ.)

Bohnet B., Nivre J., Farkas R., Ginter F., Hajic J. Joint Morphological and Syntactic Analysis for Richly Inflected Languages. *Transactions of the Association for Computational Linguistics*, 2013. Vol. 1, pp. 415–428.

Dikonov V.G., Djachenko P.V. [An experiment of building syntactic structures of English sentences using previosly known fragmentary data]. *Informatsionnye tekhnologii i sistemy (ITiS'10). Sbornik trudov 33-oi Konferentsii molodykh uchenykh i spetsialistov IPPI RAN* [Information technologies and systems (ITAS'10). Proceedings of the 33-th conference of young scientists and experts of IITP RAS]. Gelendzhik, September 18-26, 2010. Moscow, IITP Publ., 2010, pp. 310–319. (In Russ.)

Djachenko P.V., Podlesskaja O.Ju., Sizov V.G. [National corpus of Russian: the basic corpus and SynTagRus, syntactic analysis of texts with resolved morphological ambiguity]. *Informacionnye tehnologii i sistemy (ITiS'2014). Sbornik trudov 38-oi Konferentsii molodykh*

uchenykh i spetsialistov IPPI RAN [Information technologies and systems (ITAS'14). Proceedings of the 33-th conference of young scientists and experts of IITP RAS]. Nizhnii Novgorod, 2014. (In Russ.)

Iomdin L. Automatic Text Processing and Deeply Annotated Text Corpora of Russian: Interaction and Mutual Impact. *Jazykovedné štúdie XXXI.* Bratislava, Jazykovedný ústav Ľ.Štúra Slovenskej akadémie vied, 2014, pp. 136–146.

Iomdin L.L., Iomdin B.L. [The negation and valencies in Russian (according to corpus data)]. *Mezhdunarodnaya nauchnaya konferentsiya "Korpusnaya lingvistika"*, St. Petersburg, 2013, pp. 281-291. (In Russ.)

Kazennikov A.O. [A comparative analysis of machine learning dependency tree-based parsing algorithms]. *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" (2010)]. Issue 9 (16), pp. 157–162. (In Russ.)

Kazennikov A.O., Kurakin D.V. [A comparative analysis of syntactic parsing algorithms within the framework of news flow analysis]. *Trudy XVIII Vserossiiskoi nauchno-metodicheskoi konferentsii "Telematika'2011"*, St. Petersburg 2011, pp. 235–236. (In Russ.)

Kudinov M., Romanenko A., Piontkovskaya I. Conditional Random Field in Segmentation and Noun Phrase Inclination on Tasks for Russian. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2014)*. Issue 13 (20), Moscow, 2014, pp. 315–324.

Mareček D., Kljueva N. Converting Russian Treebank SynTagRus into Praguian PDT Style. *Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages 2009*, Bulgaria, pp. 26–31.

Muzychka S., Romanenko A., Piontkovskaya I. Conditional Random Field for Morphological Disambiguation in Russian. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2014)*. Issue 13 (20), Moscow, 2014, pp. 474–483.

Nivre J., Boguslavsky I., Iomdin L. Parsing the SynTagRus Treebank of Russian. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING08)*, Scott D., Uszkoreit H. (Eds.), 2008, pp. 641–648.

Podlesskaja O., Frolova T. Tagging lexical functions in Russian texts of SynTagRus. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2011).* Issue 10 (17), pp. 207–218.

Shemanaeva O.Yu. [A deeply annotated corpus of Russian texts as a teaching electronic resource]. *Informatsionnye tekhnologii i sistemy (ITiS'09).* Sbornik trudov 32-oi Konferentsii molodykh uchenykh i spetsialistov IPPI RAN [Information technologies and systems (ITAS'14). Proceedings of the 32-th conference of young scientists and experts of IITP RAS]. Bekasovo, December 15–18, 2009. Moscow, 2009, pp. 205–209. (In Russ.)

Shemanaeva O.Yu., Frolova T.I. [Tagging with lexical functions in SynTagRus]. *Informacionnye tehnologii i sistemy (ITiS'10).* Sbornik trudov 33-oj Konferencii molodyh uchenyh i specialistov IPPI RAN. Gelendzhik, 18-26 sentjabrja 2010 g. Moscow, IPPI, 2010, pp. 320–324. (In Russ.)

Täckström O., McDonald R. and Uszkoreit H. Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure. *North American Association for Computational Linguistics (NAACL)*, 2012.

Van Yun, Liu Haitao. [A quantitative study of the Russian noun according to their syntactic features]. *Vestnik Moskovskogo universiteta*. Ser. 9, Filologiya, 2013 (5), pp. 35–51. (In Russ.)