

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГУМАНИТАРНЫЙ УНИВЕРСИТЕТ»
(РГГУ)

ИНСТИТУТ ЛИНГВИСТИКИ

ФАКУЛЬТЕТ ФУНДАМЕНТАЛЬНОЙ И ПРИКЛАДНОЙ ЛИНГВИСТИКИ

УНЦ Лингвистической типологии

Павлюк Анастасия Станиславовна

**ПРЕОБРАЗОВАНИЕ СИНТАКСИЧЕСКИ АННОТИРОВАННОГО КОРПУСА
ТЕКСТОВ SYNTAGRUS В ФОРМАТ UNIVERSAL DEPENDENCIES С СОХРАНЕНИЕМ
СПЕЦИФИКИ РУССКОГО ЯЗЫКА.**

Выпускная квалификационная работа студентки 2-го курса
очной формы обучения

Направление 45.04.03 «Фундаментальная и прикладная лингвистика»

Направленность «Компьютерная лингвистика»

Допущена к защите на ГЭК

Заведующий кафедрой

Научный руководитель

(доктор филологических наук, профессор)

(доктор филологических наук, профессор)

Подлесская Вера Исааковна

Подлесская Вера Исааковна

«____»_____ 2020 г.

«____»_____ 2020 г.

Москва 2020

Оглавление

1	Введение.....	1
2	Обзор литературы и предшествующих исследований.....	3
3	Теоретический анализ синтаксических отношений.....	13
3.1	Апроксимативно-количественное СинтО.....	13
3.2	Длительное СинтО.....	17
3.3	Кратно-длительное СинтО.....	19
3.4	Дистанционное.....	20
3.5	Изъяснительное.....	20
3.6	Сравнительно-союзное СинтО.....	22
4	Практическая часть: детали конвертации.....	27
4.1	Результаты.....	28
5	Заключение.....	32
6	Список литературы.....	33

1 Введение

С момента создания Брауновского корпуса в 60-х годах прошлого столетия, корпусная лингвистика шагнула далеко вперед, расширяя как сферу применения, так и охват языков мира. Сегодня существуют корпусы множества распространенных и редких языков. Они служат не только для изучения непосредственно языка и входящих в него элементов, но и в качестве материалов для моделей машинного обучения, используемых, в свою очередь, для еще более широкого спектра применения.

Одним из современных проектов в сфере корпусной лингвистики является коллекция корпусов, объединяемая единым форматом разметки Universal dependencies (UD). Принцип, лежащий в основе данного проекта, заключается в том, чтобы размечать похожие явления в языках одинаково, при этом позволяя размечать лингвоспецифичные особенности языка. Таким образом, данный проект масштабирует свои возможности в сфере автоматической обработки текстов и позволяет проводить объемные исследования в сфере типологии языков. На момент написания данной работы в UD включены более 70 языков и более 100 корпусов, и русский язык не стал исключением в данном списке. Существует 4 русскоязычных корпуса, сконвертированных в формат UD: СинТагРус, Google Russian Treebank, Parallel Universal Dependencies, Тайга, самый крупный из которых, СинТагРус, был сконвертирован в UD в 2016 году. [5.]

СинТагРус, глубоко аннотированный корпус русского языка, был первым синтаксически аннотированным корпусом русского языка. Данный корпус имеет несколько уровней аннотации, в первую очередь морфологическую и синтаксическую разметку со снятой омонимией. Особенностью СинТагРус'а является то, что результаты автоматической разметки проверялись вручную для всей коллекции входящих в него текстов.

На данный момент объем корпуса составляет более 1,2 миллиона словоупотреблений. [15.]

Однако, соблюдение основного принципа разметки UD – размечать похожие явления в языках одинаково – требует жертв в виде лингвоспецифичных характеристик языка, которые в результате конвертации в формат UD становятся невидимыми для парсеров. В результате конвертации СинТагРус'а в формат UD 67 синтаксических отношений, размечаемых в СинТагРус'е, свели к 37 стандартным отношениям универсальной разметки. Например, кратно-длительное, дистанционное, обстоятельственно-тавтологическое, субъектно-обстоятельственное, объектно-обстоятельственное отношения в данный момент размечаются одним и тем же отношением *oblique* при том, что данное отношение как один из вариантов конвертации используется для еще 33 отношений СинТагРуса. В результате конструкции, для которых СинТагРус использует арсенал таких отношений, со всем многообразием своих значений и синтаксических особенностей, сводятся в один тип и, с точки зрения разметки, ничем не различаются. Но процесс конвертации еще не закончен, версии конвертированного СинТагРус'а обновляются, меняется логика и варианты конвертации, и сейчас есть возможность восстановить отношения и признаки, различие между которыми потерялось из-за необходимости стандартизации. Для этих целей разметка UD позволяет в дополнение к стандартизированному синтаксическому отношению указывать его лингвоспецифичный подтип.

Было принято решение снова включить в разметку следующие типы отношений: аппроксимативно-количественное, длительное, дистанционное, изъяснительное, кратно-длительное, сравнительно-союзное.

В теоретической части данной работы описано основание выбора данного списка синтаксических отношений и особенности их

функционирования. В практической части работы будет описан процесс реализации доработок конвертора.

2 Обзор литературы и предшествующих исследований

Теоретическая база разметки UD построена следующим образом: синтаксис, а именно синтаксические отношения, – это адаптированная версия системы синтаксических отношений Стэнфорда (Universal Stanford dependencies, SD) [2.], морфология - это тэги частей речи, разработанные в рамках исследований Google [9.], морфосинтаксические признаки взяты из модели конвертации разметки корпусов Interset interlingua for morphosyntactic tagsets [12.]. Тексты корпусов UD записываются в формате CoNLL-U. Данный формат был создан для оценки синтаксических парсеров на 13 языках, проводимой в рамках X конференции Conference on Computational Natural Language Learning (CoNLL) в 2006 году. [11.]

Система SD задумывалась как простая схема описания грамматических отношений в предложении. Чтобы эффективно служить целям обработки текстов, данная схема должна быть понятна не только лингвистам, но и специалистам из других сфер деятельности, которые занимаются задачей извлечения информации из текста. Поэтому при создании системы SD исследователи опирались на следующие принципы: все должно описываться в виде бинарных связей между словами в составе предложения; синтаксические отношения должны иметь четкое практическое определение, в котором, по возможности, должны использоваться традиционные грамматические понятия; то же самое касается и наименований синтаксических отношений [3.]. В итоге был создан список из приблизительно 50 грамматических отношений, которые устанавливаются между главным и зависимым словами. Число отношений разнится в зависимости от вариантов SD, которые появились из-за различных подходов

к разметке конструкций с предлогами и союзами. В частности, в разных вариантах разметки словосочетания с *союзом* и типа «*слово_1 и слово_2*» оформляются двумя способами. При первом способе одна синтаксическая связь устанавливается между двумя полнозначными словами (*слово_1 -> слово_2*), а другая связь формируется между одним из полнозначных слов (линейно первым или тем, что стоит после союза) и союзом - «*слово_2 -> и*». При втором способе союз опускается из словосочетания и превращается в разновидность синтаксического отношения, так что связь устанавливается только между полнозначными словами - «*слово_1 -> слово_2*». [1.]

Проект UD не просто перенял какие-то отдельные идеи SD, но стал по существу его продолжением в виде многоязычной версии системы синтаксических отношений Стэнфорда. В последнюю разметку UD вошли 37 синтаксических отношений SD. Что касается принципов синтаксической разметки, часть отношений перешла из SD, а часть дополнила данный список, чтобы лучше соответствовать поставленной задаче универсальности. Параллелизм в разметке разных языков был основной задачей проекта, но его разработчики понимали, что невозможно усреднить абсолютно все разнообразие явлений в языках; поэтому было решено придерживаться следующих простых, но адекватных для своих целей, идей: размечать похожие языковые явления одинаковым образом; не размечать разные языковые явления одинаковым образом, и не размечать того, чего в действительности нет в языке. В дополнение к основной разметке, разработанной на перечисленных выше принципах, была добавлена возможность размечать лингвоспецифичные явления языка. Данная возможность сохранилась и в текущей версии разметки. [3.]

Последняя версия синтаксической разметки UD устроена следующим образом [24.]: на основе синтаксических связей между главными и зависимыми словами формируется дерево зависимостей, где одно слово является вершиной всего предложения (формально оно является зависимым

служебного элемента ROOT); к данному слову также привязана вся пунктуация предложения (другими словами, пунктуационные знаки размечаются подобно зависимым словам, в поле для ссылки на главное слово у данных слов стоит ссылка на ROOT). Связи между полнозначными словами имеют приоритет над связями со служебными словами и устанавливаются напрямую между полнозначными словами, а служебные слова связываются с ближайшими к ним полнозначными словами (ср. уже приводившийся пример оформления конструкции с союзом *и*: «слово_1 -> слово_2», «слово_2 -> и»). Это означает, что служебные слова не бывают вершиной ни в каких связях. Исключение составляют лишь те случаи, когда между двумя служебными словами есть сочинительная связь, когда служебное слово управляет наречием, или, в случае эллиптических конструкций, вместо постулирования пустого элемента служебные слова в UD наделяются свойствами полнозначных слов.¹ Такое решение связано с тем фактом, что система служебных слов в разных языках различается в большой степени или может и вовсе отсутствовать, а полнозначные слова, в той или иной форме, есть во всех исследованных языках. Отдавая приоритет данным словам, мы можем соблюсти принцип одинаковой разметки схожих явлений. По той же причине вспомогательные глаголы – это всегда зависимые от смыслового глагола. В целом, можно наблюдать, что для UD характерна централизация или концентрация большей части слов предложения в позиции зависимых у меньшей части слов. Например, в конструкциях с сочинительной связью все сочиненные узлы, полнозначные слова и союзы, являются зависимыми от первого полнозначного слова. Таким образом, если сравнить деревья зависимостей, построенные на разметке UD и СинТагРус'а, первые будут содержать больше «пучков» из синтаксических отношений, а также будут короче по вертикали и шире по горизонтали. Однако нельзя не отметить, что решение разграничить служебные и полнозначные слова, а также установить между ними иерархию, по мнению некоторых исследователей, не является единственным верным, и имеет как сторонников, так и противников [8.]. Одно

из многих критических возражений состоит в том, что служебные и полнозначные слова — это не бинарная характеристика, а скорее шкала. Вовторых, по мнению критиков, данное решение лингвистически не мотивировано и идет вразрез с некоторыми принципами грамматики зависимостей. К тому же один из организаторов проекта UD, Йоаким Нивре, в своей работе 2017 года обнаружил, что при оценке парсеров, использующих разметку UD, парсеры, обрабатывающие языки аналитического строя, то есть языки с более развитой системой служебных слов, показывают более высокие результаты по сравнению с синтетическими языками, где грамматические значения массово выражаются с помощью флексий. Тем не менее, авторы работы UD поддерживают тезис о том, что иерархия между полнозначными и служебными словами способствует параллелизму в разметке языков, а для нивелирования ранее упомянутого эффекта предлагают использовать другие системы оценки эффективности, а также не использовать служебные СинтО, которые связывают служебные слова (cc, case, mark). [4.]

В UD используется не вполне общепринятое разграничение между ядерными актантами – субъектными и объектными – и другими, неядерными зависимыми вершины, в которые входят прочие актанты и сирконстанты. Для отделения ядерных актантов от остальных актантов используется следующий список критериев, действующих в разных языках [24.]:

- в большинстве языков мира глаголы-сказуемые согласуются только с ядерными актантами;
- неядерные актанты присоединяются к главному слову с помощью предлогов, в то время как ядерные присоединяются к нему беспредложно;
- некоторые падежи, традиционно называемые номинативом, аккузативом и абсолютивом, преимущественно используются для ядерных актантов;

- у ядерных актантов есть выделенная позиция в предложении, чаще всего рядом с глаголом;
- только ядерные актанты могут контролировать вершину зависимых клауз с невыраженным подлежащим, а также образовывать придаточные относительные.

Решение о подобном разграничении актантов было принято исходя из того факта, что в большинстве языков существует (немаркированный) способ присоединения одного или двух аргументов к глаголу (в зависимости от типа глагола – переходный, непереходный) без дополнительных служебных слов.

Кроме разграничения между актантами с помощью синтаксических отношений, в UD отделяют субъекты от комплементов, связь идущую к предикату, от связи, идущей к именными группами, конструкции с обязательным контролем от других способов введения подлежащего (в этих целях используется отношение `xcomp`: open clausal complement; данное отношение устанавливается между предикатами главной и зависимой клаузы), а также отличают вложенные клаузы, характеризующие существительное, но при этом не являющиеся определительным придаточным. [24.]

Корпус СинТагРус является составной частью Национального корпуса русского языка и построен на принципах лингвистического процессора ЭТАП-3. Именно с помощью парсера этого процессора выполняется первичная разметка текстов. Процессор ЭТАП-3 в свою очередь основан на теории «Смысл-Текст» И. А. Мельчука, которая также представляет синтаксическую структуру предложения в виде дерева зависимостей [14.]. Примерный список синтаксических отношений в рамках теории «Смысл-Текст» использовался для «системы поверхностно-синтаксического анализа» и состоял из 42 отношений [21.]. Данный список составлен на материале научно-технических текстов и подходит для деловой прозы нейтрального стиля. Поэтому предполагалось, что список поверхностно синтаксических

отношений будет дорабатываться, что и было сделано в синтаксическом парсере ЭТАП-3 и применено в рамках проекта СинТагРус. Итоговая не конвертированная версия корпуса содержит 67 отношений, которые относительно универсальны для всех стилей и жанров, хотя тексты корпуса преимущественно относятся к публицистическому, научно-популярному, художественному и новостному жанрам. [25.]

Важной особенностью корпуса СинТагРус является процесс обработки текстов. Текст, разделенный на предложения, пропускается через парсер процессора ЭТАП-3. В результате создается дерево зависимостей, в узлах которого находятся слова предложения с описанием своих морфосинтаксических свойств, а ветви помечены наименованиями синтаксических отношений. На этом же этапе происходит автоматическое снятие омонимии. Следующим этапом является ручная проверка разметки лингвистами. СинТагРус является единственным крупным полностью отредактированным синтаксическим корпусом русского языка [14.]. Среди других отличий СинТагРус'а имеется ряд решений в сфере разметки, а именно, в синтаксической структуре эллиптических предложений восстанавливаются опущенные слова с соответствующими контексту морфологическими характеристиками; многословные лексические единицы либо представляются как один узел, либо связываются вспомогательным отношением, устанавливаемым между элементами таких единиц; прямая речь также размечается двумя способами: с помощью комплетивного отношения между вершинами авторских слов и прямой речи, и с помощью вводного отношения, если предложение начинается с прямой речи. С точки зрения служебных слов, предлогов и союзов, также есть свои особенности. В отличии от UD предлоги в СинТагРус'е могут быть вершинами, а связь от предлога к полнозначному слову выражается через предложную связь. [25.]

С точки зрения более общих принципов UD и СинТагРус похожи в выборе приоритетов для решений, связанных с описанием синтаксиса. Так

один из организаторов проекта UD пишет, что, несмотря на то, что все решения должны строиться на имеющейся парадигме лингвистической теории, порой приходится идти на компромисс ради практической ценности и удобства полученной в итоге разметки [3.]. Также и И. А. Мельчук пишет, что формальное описание синтаксической структуры должно «разумно стремиться к максимальной формальной простоте применяемого описания» [23.]. Тем не менее, очевидно, что требования формальной простоты и удобства применения могут быть выполнены разными способами. Кроме общей основы в виде грамматики зависимостей и отчасти совпадающего набора частей речи и морфосинтаксических характеристик, между рассматриваемыми разметками существует множество различий, требующих решений при конвертации. Поэтому разработчики первой версии конвертора СинТагРус 'а выработали алгоритм, состоящий из 4 этапов [5.]. На первом этапе предлоги и союзы, являющиеся вершинами в СинТагРус 'е, преобразовали в зависимые, как это принято в UD.

Второй этап включал конвертацию синтаксических отношений. Для этого структура деревьев была преобразована согласно вербоцентричной разметке UD, которая наделяет приоритетом полнозначные слова. Для преобразования самих отношений были разработаны правила. В первую очередь по списку соответствий преобразовывались отношения, имеющие одно соответствие в UD. Затем, согласно разработанным правилам, преобразовывались отношения, которые могут иметь два и более соответствий, или те соответствия, которые невозможно установить, опираясь только на оригинальное синтаксическое отношение. Правила для данных случаев опираются на части речи вершин и зависимых, а также их морфосинтаксические характеристики. Пример простого правила: отношение «предик» конвертируется в «nsubjpass», если слово или вершина данного слова имеет признак «СТРАД», иначе используется отношение «nsubj».

Примером более сложного правила является конвертация сравнительного отношения в «nmod», «advmod» или «advcl».

Третьим этапом происходила конвертация морфологической разметки. Аналогично второму этапу для этого применялись либо списки соответствий между маркировками СинТагРус ’а, либо более развернутые правила, что, в случае морфологии, требовалось реже.

Четвертый этап включал преобразование формата данных. Тексты СинТагРус ’а хранятся в XML-файлах по стандарту специализированного языка разметки TEI (Text Encoding Initiative) [27.], что потребовало конвертации в формат CoNLL-U. Также на данном этапе переносилась пунктуация. [5]

Первая версия конвертора использовала ограниченное число лингвоспецифичных помет, однако большее их количество планировалось реализовать в будущих версиях. Конвертор дорабатывался, а версии СинТагРус ’а в UD обновлялись, в том числе в рамках общего обновления версий UD. Среди прочих изменений в рамках перехода ко второй версии UD добавилось различие между именными определениями существительного и предиката – в первой версии UD и то, и другое размечалось отношением «nmod», во второй версии данное отношение используется только для именных определений существительного, определения предиката размечается отношением «obl». Также во второй версии отказались от отдельных отношений «nsubjpass», «csubjpass» and «auxpass». Вместо этого в языках, где различие между пассивным и активным залогом является существенным, к отношениям «nsubj», «csubj» and «aux» стали добавлять лингвоспецифичные пометы «:pass». Таким образом, короткий список лингвоспецифичных помет первой версии конвертора, состоящий из одного отношения «acl:relcl» для определительного придаточного, дополнился следующими наименованиями: «nsubj:pass» для именных субъектов глагола в пассивном залоге, «csubj:pass» для клаузы, несущей роль субъекта,

«flat:name» для связи имени собственного, состоящего из нескольких слов. В таблице ниже можно увидеть список всех соответствий конвертированных отношений, а также разницу между первой и последней версией разметки.

Таблица 1.1. Соответствия СинтО в СинTagРус'е и UD

Отношение в СинTagРус'е	сокращение	Отношение в UD (первая версия)	Отношение в UD (последняя версия)
Описательно-определительное	оп-опред	acl	amod/ acl
Субъектно-кодикативное	суб-копр	acl	obl/ nmod/ advmod/ advcl/ acl
Объектно-кодикативное	об-копр	acl	obl/ advmod/ nmod/ advcl/ xcomp
Эксплитивное	эксплет	acl	advmod/ nmod/ advcl/ acl/ obl/ mark
Релятивное	релят	acl:relcl	acl:relcl
Подчинительно-союзное	подч-союзн	advcl	advcl/ nmod/ obl/ amod/ acl/ parataxis/ advcl/ acl:relcl/ appos/ ccomp/ discourse/ mark
Инфинитивно-союзное	инф-союзн	advcl	advcl
Ограничительное	огранич	advmod	nmod/ advcl/ acl/ fixed/ advmod/ discourse/ aux/ expl/ mark/ obl/ dep
Количественно-ограничительное	колич-огран	advmod/nmod	nmod/ obl/ advmod
Обстоятельственное	обст	advmod/nmod/ advcl/acl/xcomp	obl/ advmod/ nmod/ advcl/ mark/ xcomp
Определительное	опред	amod	amod/ nmod/ acl
Комплетивно-аппозитивное	компл-аппоз	amod/nummod/ nmod/acl	appos/ nummod/ nmod/ obl/ advmod
Кратное	кратн	amod/nummod/ nmod/advmod/xcomp	xcomp/ advmod/ amod/ obl/ flat/ conj/ fixed/ flat:name/ nmod/ /nummod
Обособленно-аппозитивное	об-аппоз	appos	appos
Пролептическое	пролепт	appos/cop (это, вот)	appos/ cop (это/ вот)
Аппозитивное	аппоз	appos/name	appos/ parataxis
Номинативно-аппозитивное	ном-аппоз	appos/nmod	appos
Примыкальное	примыкат	appos/parataxis	nmod/ acl/ parataxis/ advmod/ advcl/ obl/ appos/ flat:name
Аналитическое	аналит	aux/auxpass	aux/auxpass
Пассивно-аналитическое	пасс-анал	auxpass	aux:pass
Предложное	предл	case/mark	fixed
Сочинительное	сочин	cc/conj	cc/conj
Сентенциально-сочинительное	сент-соч	cc/conj	
Сравнительное	сравнит	cc/nmod/advmod/ advcl	mark/ cc/ nmod/ amod/ obl/ acl/ advmod/advcl/
Соотносительное	соотнос	cc/nmod/advmod/ nummod	obl/ advcl/ mark/ cc/ nmod/ amod/ fixed
Композитное	композ	compound	compound

Количественно-вспомогательное	колич-вспом	compound	compound
Сочинительно-союзное	соч-союзн	conj/cc	conj/cc
Присвязочное	присвяз	cop	xcomp/ obl/ advmod/nmod/ advcl
1-е комплективное	1-компл	dobj/iobj/ccomp/ xcomp	advcl/ acl/ ccomp/ obl/ parataxis/ nmod/ discourse/ xcomp/ appos/ obj
2-е комплективное	2-компл	dobj/iobj/ccomp/ xcomp	obl/ nmod/ advcl/ acl/ xcomp/ appos/ iobj/ obj
3-е комплективное	3-компл	dobj/iobj/ccomp/ xcomp	obl/ advmod/ nmod/ advcl/ xcomp/ acl/ appos/ iobj/ obj
4-е комплективное	4-компл	dobj/iobj/ccomp/ xcomp	obl/ advmod/ nmod/ advcl/ xcomp/ acl/ appos/ iobj/ obj
5-е комплективное	5-компл	dobj/iobj/ccomp/ xcomp	iobj/ccomp/xcomp/ obl/ acl/ appos/ iobj/ obj/ nmod
Сентенциально-предикативное	сент-предик	expl	expl
Дательно-субъектное	дат-субъект	iobj/nmod	iobj/ nmod/ obl
Неактантно-комплективное	неакт-компл	iobj/nmod	iobj/ nmod/ obl
Вспомогательное	вспом	name/mwe	fixed/ flat:name/ obl/ discourse
Квазиагентивное	квазиагент	nmod	nmod/ advmod/ obl
Несобственно-агентивное	несобст-агент	nmod	iobj/ nmod/ obl
Элективное	электив	nmod	nmod
Адресатно-присвязочное	адр-присв	nmod	nmod
Атрибутивное	атриб	nmod	Nmod/ advmod/ obl/ acl/ advcl/ flat
Распределительное	распред	nmod	nmod/ obl
AUDитивное	aUDит	nmod	
Кратно-длительное	кратно-длительн	nmod	obl
Дистанционное	дистанц	nmod	obl
Обстоятельственно-тавтологическое	обст-тавт	nmod	obl
Субъектно-обстоятельственное	суб-обст	nmod	obl
Объектно-обстоятельственное	об-обст	nmod	obl
Длительное	длительн	nmod/advcl	advmod/ obl/ nmod
Сравнительно-союзное	сравн-союzn	nmod/advcl/ advmod	advcl/ acl/ nmod/ advmod/ obl
Уточнительное	уточн	nmod/advmod/cc	nmod/advmod/ obl acl
1-е несобственно-комплективное	1-несобст-компл	nmod/ccomp/ xcomp	Xcomp/ ccomp/ nmod/ advmod/ obl
2-е несобственно-комплективное	2-несобст-компл	nmod/ccomp/ xcomp	Xcomp/ ccomp/ nmod/ advmod/ obl
3-е несобственно-комплективное	3-несобст-компл	nmod/ccomp/ xcomp	Xcomp/ ccomp/ nmod/ advmod/ obl
Агентивное	агент	nmod:agent	obl / nmod
Коммуникативно-сочинительное	ком-сочин	no data with elliptical	
Предикативное	предик	nsubjpass/nsubj	nsubj/csubj/nsubj:pass/csubj:pass
Количественное	количест	nummod	nummod

Аппроксимативно-количественное	аппрокс-клич	nummod	nummod
Количественно-копредикативное	клич-копред	nummod/nmod/advmod	Advmod/ obl/ advcl
Нумеративно-аппозитивное	нум-аппоз	nummod:appos	nummod:entity
Вводное	вводн	parataxis	parataxis/ vocative
Изъяснительное	изъясн	parataxis	parataxis
Разъяснительное	разъяснит	parataxis	parataxis
Леммы: “МИЛЛИАРД”, “МИЛЛИОН”, “ТРИЛЛИОН”, “ТЫСЯЧА”, “БИЛЛИОН”.		“nummod:gov”	nmod

Важным отличием последней версии, с точки зрения русского корпуса UD, является разметка эллипса. В первой версии конвертора эллиптические конструкции, которые в сумме составляли 2800 предложений, никак не обрабатывались. Теперь же используется следующая схема: роль опущенного слова выполняет одно из его зависимых слов, а в случаях, когда такая замена невозможна из-за образования аномальных связей между словами, используется специальное отношение «*orphan*».

3 Теоретический анализ синтаксических отношений

Данная глава содержит анализ синтаксических отношений СинтагРус'а, подлежащих переносу в разметку UD в виде дополнительных лингвоспецифичных помет, а также основания для выбора данных отношений.

3.1 Аппроксимативно-количественное СинтО

Аппроксимативно-количественное СинтО (аппрокс-клич) используется для представления конструкций со значением приблизительного количества. Вершина (X) в данных конструкциях – существительное или его функциональный аналог, а зависимое (Y) – числительное в постпозиции. Примеры: *минут [X] пять [Y]*; *Они все должны быть в море часов через 12 после крушения [25.]*. Таким образом, с помощью

порядка слов выражается субъективно-модальное значение «приблизительности и неуверенности». Выражения со значением субъективной модальности передают отношение говорящего к сообщаемому и могут быть экспрессивно окрашены, что накладывает ряд ограничений на их использование в разных стилях [30.]. Рассматриваемая конструкция практически, не используется в официально-деловом и научном стилях, но распространена в разговорном и художественном [20.]. По мнению некоторых лингвистов, степень экспрессивности зависит от языкового уровня использованных средств [26.]. Так, например, лексические средства более нейтральные, грамматические же, к которым относится и порядок слов, являются более экспрессивными, что объясняет подобное распределение использования аппроксимативно-количественного отношения по функциональным стилям.

Конструкции с аппроксимативно-количественным СинтО имеют свои ограничения на вид числительных в постпозиции. В данных конструкциях не могут употребляться числительные со спорным статусом - слова со значением числительного и морфологическими признаками существительного, среди которых [22.]:

- числа, кратные 1000 (*рублей две тысячи», *роз миллион);
- «целая» (*процентов одна целая и три десятых);
- «половина» (*коробок половина)
- дробные доли (*килограмма треть)
- «нуль» и «ноль» (вариантов ноль – допустимо однако в данном случае это полное высказывание, оформляющееся не аппроксимативно-количественным СинтО).
- *Один* (*пришло человека два*, но не **пришел человек один* и даже не **пришел человек двадцать один*).

В аппроксимативно-количественной конструкции допускается наличие между вершиной и зависимым других слов, однако их количество

ограничено. Самая распространенная часть речи между участниками конструкции это предлоги, другие менее распространены. В части СинТагРус'а входящей в НКРЯ на 196 вхождений с аппроксимативно-количественной связью приходится всего 3 предложения, где между существительным и числительным присутствует более одного слова: «Он находится ровно над нами, метрах так в двадцати.»; «Но вот мы встретимся однажды, совсем уже не в наше время, лет почти через двадцать; «Правда, и сейчас в его наряде можно наблюдать некоторую искреннюю задержку года, скажем, на три». Из результатов видно, что кроме предлогов допустимы наречия и вводные конструкции. Однако не каждый контекст позволяет встраивать дополнительные слова между вершиной и зависимым, в частности конкретизаторы [16.]. Например: *Он съел штук (*уже) пять этих десертов.*

Что касается семантики используемого числительного, в аппроксимативных конструкциях в основном используются круглые числа. Более того, использование «некруглых» чисел значительно ограничено, т. к. это противоречит значению приблизительности [17.]. Однако данное ограничение не является строгим, ведь математическое округление может происходить до разных разрядов. Поэтому можно предположить, что в сферах деятельности, где принято вести исчисления в десятых, сотых числа и меньше, в разговорной речи допустимо использовать относительно точные числа в рассматриваемых конструкциях, в силу того, что в данном контексте они будут «круглыми». Пример: *Температура в криостате повысилась процента на три с половиной.* Статистика зависимых слов в аппроксимативных конструкциях корпуса подтверждает распространенность круглых чисел и допустимость более точных чисел. (см. Таблица 3.2¹).

¹ В таблице подсчитана также частота числительных связанных с непосредственными зависимыми в конструкциях с аппроксимативно-количественным СинТО кратным и количественно-вспомогательным отношениями, а также союзом «или» т.к. данные отношение и союз, по сути, уравнивают их и распространяют на них значение приблизительности.

Таблица 3.2. Частота лексем зависимых слов в аппроксимативно-количественных конструкциях

Зависимое	Частота
Десять	39
Двадцать	25
Два	24
Три	20
Пять	19
Пятнадцать	13
Шесть	10
Тридцать	10
Четыре	7
Семь	6
Сто	6
Полтора	5
Сорок	5
Семьдесят	5
Двендцать	4
Пятьдесят	3
Четырнадцать	3
Восемьдесят	3
Двадцать пять	2
Двести	2
Четыреста	2
Пятьсот	2
Восемь	2
Шестьдесят	2
Тринадцать	2
Триста	1
Одиннадцать	1
Сто пятьдесят	1
Двадцать семь	1

В выборке из 1154 языков в 479 числительное предшествует существительному, в 608 числительное находится в постпозиции, и в 65 языках допустимы оба порядка членов числовой конструкции, и ни один из них не превалирует [7.]. При этом русский язык относится к первому типу [10.], т.к. порядок «числ. + сущ.» является нейтральным и самым распространенным способом выражения количества чего-либо, аппроксимативные же конструкции являются маркированными и помимо количества обладают дополнительным значением, описанным выше. Тем не менее, невозможно отрицать, что в русском языке инвертированные числительные конструкции допустимы и являются его особенностью. По этой причине русский язык можно сравнивать с языками третьего типа, где

допустимы оба порядка, при этом выбор порядка влияет на значение числительной конструкции. Например, в языке ниас позиция числительного является маркером категории определенности всего словосочетания: постпозиция числительного – определенное, наборот – неопределенное. Пример:

(4) Nias (Brown 2005)			
a.	öfa	geu	m-baβi s=afusi
	four	CLF	ABS-pig REL=white
			'four white pigs'
b.	baβi-ra	s=afusi	si=öfa geu
	pig-3PL.POSS	REL=white	REL=four CLF
			'their four white pigs'

Рисунок 1. Порядок слов в числительных конструкциях и категория определенности в языке ниас (источник: WALS)

На данный момент количество вхождений аппроксимативно-количественного отношения в СинтагРус 'е равно 196. В UD данное СинтО преобразуется в отношение nummod так же, как и количественное СинтO, и таким образом ничем не отличается от него с точки зрения парсера.

3.2 Длительное СинтO

Длительное СинтO устанавливается между главным словом – вершиной предложения и зависимым – обстоятельством длительности со значением «приблизительного количества или распределительности», выражаемым существительным в винительном падеже или эквивалентной ему предложной группой. Пример: *Велосипед, например, с которого он не слезал [X] три года [Y] [25.]*. Однако данное определение стоит дополнить исходя из примеров корпуса. Вершиной может быть не только глагол, но и другие части речи, выполняющие роль предиката. Например, существительное: «Все остальное время [Y] ее основная задача [X] в доме – отлично выучить детей...»; предложная группа: «В большинстве случаев

археологи всё время [Y] в [X] поиске денег»; наречие: «...после нашей встречи у него съемки на телевидении, затем снова интервью и встречи – и так [X] две недели [Y] подряд»; прилагательное: «Между тем, уже более сорока лет [Y] известно [X], что...». Ввиду того, что в СинТагРус 'е местоимения не рассматриваются как отдельная часть речи и, в зависимости от морфологических и синтаксических свойств, размечаются как существительные, прилагательные или наречия, примерами длительного СинтO являются и следующие предложения: «Палец плавно нажимал на спуск, проходя тот отмеренный срок, который [X] еще оставалось жить [X] немцу», «За тот месяц, что [Y] оностоял [X], была собрана солидная база по кредитным картам». Кроме дополнения определения в части вариантов главного слова, стоит отметить, что зависимое слово также может быть не только существительным или предложной группой, но и наречием (см. Таблица 3 .3). Например, «Большинство банков, у которых в последнее время была отозвана лицензия, имели отрицательный капитал более двух лет.».

Следует отметить, что в СинТагРус 'е фразеологизмы вроде «из года в год», или «с утра до вечера», или словосочетания вроде «с 10 до 14 лет», которые также выражают длительность, не размечаются длительным СинтO, что в ряде случаев представляется непоследовательным. С вершиной предложения они связываются обстоятельственным отношением, а внутри конструкции для парных предлогов используется соотносительное СинтO. Так как при конвертации и длительное, и обстоятельственное отношения (в подобных конструкциях) преобразуются в одно отношение «obl», данное различие в разметке конструкций со значением длительности нивелируется. Однако в рамках данной работы лингвоспецифичная помета, как для длительного отношения, добавлять не будет.

Длительное СинтO является достаточно распространенным отношением и насчитывает в корпусе 778 вхождений. В UD СинтO

преобразуется в одно из следующих отношений в зависимости от частей речи главного и зависимого слова:

Таблица 3.3. Соответствия частей речи участников конструкции и выбора СинтО из UD

Часть речи главного слова	Часть речи зависимого слова	СинтО в UD	Частота
Глагол	Наречие	advmod	72
Глагол	Существительное	obl	640
Существительное	Существительное	nmod	4
Наречие	Существительное	obl	2
Прилагательное	Существительное	obl	8
Глагол	Прилагательное	obl	2

3.3 Кратно-длительное СинтО

Кратно-длительное СинтО также используется для передачи значения длительности. Вершина (X) в данных конструкциях – это глагол, зависимое (Y) – обстоятельство многократной длительности в творительном падеже и множественном числе, что отличает данное отношение от длительного. Значение «многократной длительности» реализуется благодаря множественному числу слов со значением длительности. Пример: *Там Михаил Сергеевич собирал республиканских лидеров и часами [Y] что-то говорил [X] [25.]*. (такое событие повторялось, оно не могло быть единичным).

Среди семантических ограничений на участников кратно-длительного отношения следует отметить невозможность использовать в качестве обстоятельства длительности слова, обозначающие относительно короткие промежутки времени, например, «секундами», «минутами». Пример: *Там Михаил Сергеевич собирал республиканских лидеров и *секундами что-то говорил.* Это объясняется тем, что кратно-длительная конструкция характеризует не любую, а большую продолжительность события. Поэтому возможно *Он целыми часами пропадал в гараже*, но не **Он всего часами пропадал в гараже*.

В корпусе содержится 65 вхождений кратно-длительного СинтО. Данное СинтО является одним из многих СинтО СинТагРус'а преобразуемых в отношение *oblique*.

3.4 Дистанционное

Еще одним отношением, характеризующим глагол количественными категориями, является дистанционное отношение, в котором, соответственно, вершина – глагол, зависимое – обстоятельство пространственной протяженности, выражаемое существительным в винительном падеже или предложной группой со значением приблизительного количества или распределительности. Пример: *Купив на станции в Пушкино огромный арбуз, он все три километра [Y] до нашей дачи катил [X] его перед собой по дороге* (Ю.М.Нагибин) [25.].

В конструкциях с беспредложным примыканием винительного падежа реализуется значение меры. В отличие от меры времени (длительное СинтО), в конструкциях передающих меру пространства на глагол накладываются лексико-семантические ограничения, т. к. любое действие может длиться некоторое количество времени, но не каждое может иметь пространственную протяженность [29.]. Поэтому вершиной конструкций с дистанционным отношением могут быть глаголы движения.

Количество вхождений дистанционного отношения относительно невелико – всего 5 вхождений на весь корпус. В результате конвертации СинтО преобразуется в отношение *oblique*.

3.5 Изъяснительное

В изъяснительном СинтО вершина – это вершина группы сказуемого главного предложения, а зависимое – вершина придаточного, включающего союзное слово «что», «отчего» или «почему». Придаточное, введенное данными союзными словами, характеризует и раскрывает содержимое

главного предложения. Пример: *Они решили разогнать [Х] толпу, что и было [Y] сделано.* [25.]

Несмотря на то, что изъяснительное СинтО СинТагРус 'а используется для связи сложноподчинительного придаточного, оно не является идентичным другому традиционному термину «изъяснительное придаточное» [19.]. Чаще можно найти определение [28.], что изъяснительная связь между главным и придаточным предложением служит для связи придаточных, отвечающих на вопросы косвенных падежей и относящихся к члену главного предложения, который нуждается в смысловом расширении (у Н. Ю. Шведовой. – информативно-недостаточный глагол [29.]). Примеры опорного слова: «грустить», «грустно», «удивляться», «удивлен», «удивление», «говорят», «слышать», «слышно», «хорошо». Несмотря на то, что изъяснительное отношение, как и все СинтО в СинТагРус 'е формируют конструкции, состоящие из одного главного слова и зависимого, его смысловым отличием от вышеупомянутых изъяснительных придаточных видится в опоре на все главное предложение, а не на отдельное слова. Изъяснительное придаточное по Н. Ю. Шведовой имеет конкретное опорное слово в главном предложении и в зависимости от синтаксических и семантических характеристик данного слова и связующего служебного слова, характеризует, изъясняет именно опорное слово. Пример: «Сегодня он заметил, что график изменился» (в СинТагРус 'е такие конструкции оформляются в основном комплетивными СинтО). В данном примере придаточное относится к одному слову из главного предложения – «заметил», и является его смысловым распространением. Изъяснительное же придаточное СинТагРус 'а с точки зрения смысла, а не формальной структуры СинтО, ссылается на все главное предложение, т.к. базируется на анафорическом употреблении местоименного существительного и наречия («что», «отчего», «почему»), для которых все главное предложение является антецедентом [29.]. Пример: «Он заметил изменения в графике дежурств,

что и вызвало его подозрения. В данном примере невозможно выделить одно слово главного предложения, которое может быть антецедентом союзного слова «что», т. к. в придаточном характеризуется содержимое всего главного предложения.

Союзные слова «отчего» и «почему» ссылаются на главное предложение, как на причину того, о чём идет речь в придаточном предложении. Следствием данного факта является то, что придаточные предложения, введенные этими союзными словами, должны находиться в постпозиции по отношению к главному [30.]. Пример: **Отчего содержание сахара в крови резко падает, этот гормон направляет глюкозу в печень и мышцы*. В отношении союзного слова «что» данное ограничение не действует. Пример: *«Что всегда интересовало киевскую власть, так это крымская земля, особенно на Южном берегу»*.

В корпусе содержится 604 вхождения с изъяснительным отношением. В результате конвертации СинтО преобразуется в отношение parataxis также, как и разъяснительное, а также 5 других СинтО, где parataxis используется как вариант конвертации.

3.6 Сравнительно-союзное СинтО

Как уже было отмечено в обзоре литературы, одним из основных отличий между UD и СинтагРус'ом является разметка служебных и полнозначных слов. В UD связи между полнозначными словами имеют приоритет над связями со служебными словами и устанавливаются напрямую между ними, служебные же слова связываются с ближайшим полнозначным словом с помощью отношений «mark», «cc» или «case». В СинтагРус'е с точки зрения роли в синтаксических отношениях полнозначные и служебные слова равны по статусу, и служебные слова могут быть главными словами. Например, конструкции, где два полнозначных слова объединены союзом, оформляются следующим образом:

одно СинтО устанавливаются между первым полнозначным словом и союзом, второе между союзом и вторым словом. Данный факт позволяет выделить конструкции со сравнительными союзами, которые в UD конвертируются в СинтО case. Сравнительные союзы СинтагРус'е связываются с помощью двух отношений – сравнительного и сравнительно-союзного.

Сравнительно-союзное устанавливается между главным словом – сравнительным союзом и зависимым – вторым из сравниваемых членов сравнительной конструкции. Например, «*Детские вопросы, как [X] правило [Y], самые трудные*» [25.]. В СинтагРус'е насчитывается 5 166 вхождений с данным отношением.

Сравнительное СинтО используется для двух типов конструкций. Главное слово – это всегда прилагательное или наречие в сравнительной степени, а зависимое слово может быть именной группой в родительном падеже, являющейся второй из сравниваемых членов, или сравнительным союзом. Например, «Это не груз для корабля таких размеров, в нем одной рыбы было немногим меньше [X] полутора тысяч [Y] тонн!», «Нам гораздо важнее [X] сотрудничество, чем [Y] отражение каких-то явных и неявных угроз» [25.]. Всего в корпусе 4 408 подобных конструкций.

При конвертации из-за разницы приоритетов полнозначных и служебных слов полнозначное слово, являющееся зависимым союзом в СинтагРус'е, занимает его место, т.е. данное слово перенимает СинтО союза, а также становится зависимым полнозначного слова, к которому был привязан союз. Союзу при этом достается роль зависимого этого слова («mark», «cc», «case»). Например, конструкция «... их рассматривали как источник бед» («Но акции устрашения зачастую производились в ответ на действия партизан и подпольщиков, которые вызывали у населения неоднозначную реакцию: их рассматривали как источник бед и несчастий.»),

имеющая в СинТагРус'е следующую структуру, на промежуточном этапе конвертации выглядит следующим образом:

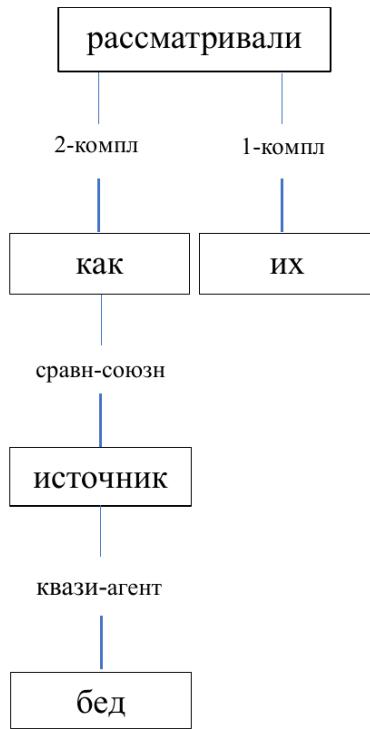


Рисунок 2. Оригинальная конструкция



Рисунок 3. Переходная конструкция

Таким образом сравнительное и сравнительно-союзное СинтО преобразуется во множество различных отношений (см. Таблица 3.4), затем в зависимости от морфосинтаксических характеристик данного зависимого

слова и его нового главного слова между ними устанавливаются отношения «nmod», «amod», «acl», «obl», «advcl», или «advmmod». Однако наличие пометы case, используемой для связи сравнительных союзов с полнозначными словами, избавляет от необходимости добавлять лингвоспецифичную помету к каждому СинтО из данного множества. Для выделения необходимых конструкций будет использована только одна синтаксическая помета case:compar.

Таблица 3.4. Частота СинтО, наследуемых от союза

№	СинтО наследуемое зависимым словом от союза	Частота
1.	Сравнит	2 989
2.	Вводн	1 109
3.	2 компл	324
4.	Соотнос	100
5.	Эксплэт	96
6.	Сочин-союзн	73
7.	1 компл	67
8.	Примыкат	55
9.	Разъяснит	28
10.	3 компл	25
11.	Обст	22
12.	Присвяз	22
13.	Сочин	22
14.	сент-соч	13
15.	об-копр	12
16.	подч-союзн	10
17.	суб-копр	10
18.	предик	8
19.	предл	7
20.	4 компл	4
21.	Релят	4
22.	Огранич	3
23.	Атриб	3
24.	Аппоз	1
25.	Сравн союзн	1

Из статистика наследуемых СинтО видно, что чаще всего бывшее зависимое союза в сравнительно-союзных конструкциях перенимает сравнительное и вводное СинтО.

Русский язык не единственный в своей группе родственных славянских языков, в разметке которого в UD есть лингвоспецифичная помета для

сравнительных конструкций. В польском корпусе используются следующие СинтO: «advcl:cmpr» « (comparative clause), «obl:cmpr» (comparative phrase). [24.]

Отношение «advcl:cmpr» используется для предикативных сравнительных конструкций.

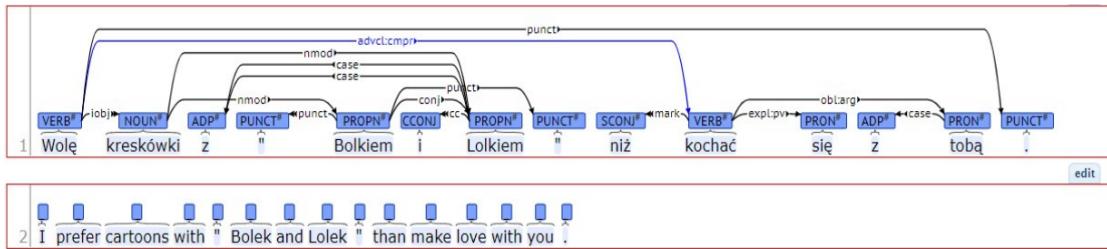


Рисунок 4. Пример СинтO «advcl:cmpr» в польском языке

Отношение «obl:cmpr» используется для сравнительных конструкций без предиката.

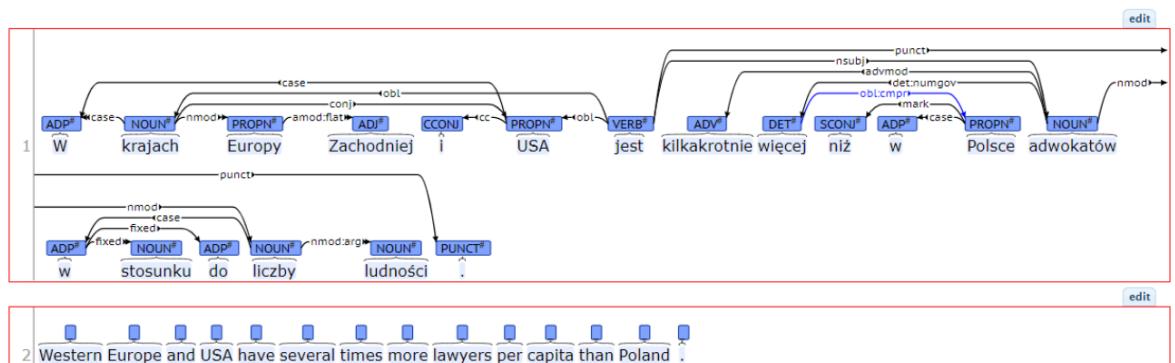


Рисунок 5. Пример СинтO «advcl:cmpr» в польском языке

В русском языке сравнительные конструкции также бывают с предикатом – придаточные сравнительные – и без него – сравнительные обороты. Однако строгое разделение данных типов так, как это реализовано в польском, не совсем подходит для русского языка, т.к. в нем еще существуют неполные придаточные сравнительные, где предикат совпадает с предикатом главного предложения и потому может опускаться, но тем не менее данный опущенный предикат обладает собственными зависимыми словами [30.].

Например, «Ямщик был в таком же изумлении от его щедрости, как и сам француз от предложения Дубровского».

4 Практическая часть: детали конвертации

Конвертор СинТагРус'а состоит из 16 модулей, каждый из которых выполняет определенную часть обработки. Данные модули объединяются и запускаются файлом, при этом запускается своего рода конвейер обработки: на вход первому модулю подаются материалы СинТагРус'а, результат работы модуля сохраняется в отдельную папку, входными данными следующего модуля будут файлы из данной папки, и он также сохранит результат в отдельной папке. Данный процесс продолжается таким же образом до последнего модуля, на выходе которого будут файлы корпуса с разметкой UD.

Синтаксические отношения обрабатываются отдельным синтаксическим модулем. На момент запуска данного файла согласно стандартам UD уже исправлена разметка составных слов, предлогов, союзов и числительных. Конвертация синтаксических отношений, как уже было описано в главе теоретического обзора литературы, строится из 3 типов правил. При этом в рамках синтаксического модуля сначала выполняется обработка в соответствии со сложными правилами, затем односложные правила, где замена СинтО происходит один к одному, затем правила, основанные на части речи главного и зависимого слова. Большая часть отношений, выбранных для добавления в разметку UD в качестве лингвопсемиотических помет, имеет одно соответствие в UD, следовательно, добавление данных помет, так же, как и для конвертации отношений один к одному, не требует сложного алгоритма. Исправление кода в данных случаях заключается в добавлении помет к названиям итоговых СинтО.

Для разметки текста синтаксическими отношениями в UD существуют пометы первого и второго уровня. Первый уровень – это стандартный список

универсальных помет, второй – лингвоспецифичные пометы. Пометы обоих типов находятся в колонке DEPREL и разделяются символом «::». DEPREL – это седьмая из десяти позиций разметки, в которых хранится вся информация о слове в составе предложения. [5.]

В результате были добавлены следующие пометы второго уровня:

Таблица 4.5. Пометы второго уровня

Отношение в СинТарРус’е	Отношение в UD (последняя версия)	Новый тег
Кратно-длительное	obl	obl:mdur
Дистанционное	obl	obl:dist
Длительное	advmod/ obl/ nmod	advmod:dur / obl:dur/ nmod:dur
Аппроксимативно-количествоное	nummod	nummod:approx
Изъяснительное	parataxis	parataxis:relcl
Сравнительно-союзное (связь с союзом)	case	case:compar

Кроме изменений в синтаксисе правки также вносились в файл обработки союзов, в котором обрабатываются случаи, когда союз является вершиной предложения. Среди конструкций со сравнительно-союзовым отношением насчитывается 158 вхождений, где главное слово, союз, является вершиной предложения. Поэтому в алгоритм был добавлен словарь сравнительных союзов, а также дополнительная проверка на данный тип союза. По результатам проверок союзы и союзные слова конвертируются в одно из 3 отношений: сравнительные союзы в «case:compar», остальные придаточные союзы в «mark», сочинительные в «cc».

4.1 Результаты

В результате реализации списка лингвоспецифичных помет, являющихся предметом данной работы, в русском корпусе UD можно выделить числительные конструкции со значением приблизительности. Ранее предложения из примера «см. Таблица 4 .6» не отличались от

немаркированных количественных конструкций «EX. 2.» с точки зрения парсера.

Таблица 4.6. Примеры 1

	СинTagРус'	UD	Пример
EX. 1.	аппрокс- коляч	nummod:approx	Уже лет [X] семь [Y] я занимаюсь бытовой семантикой и всё равно бокалы называю рюмками.
EX. 2.	количество	nummod	Он разъезжал по своим раскопкам в высокой коляске, а за ним следовали трое [Y] его помощников [X] на велосипедах

В русском корпусе UD также используется отношение «nummod:gov» для числительных конструкций с леммами «миллиард», «миллион», «триллион», «тысяча» and «бillion». Однако в случае конвертации аппроксимативно-количественного отношения, при добавлении новой пометы второго уровня «:approx» нет конфликтов с пометой «gov», т. к., как уже было упомянуто в теоретической части, в данных конструкциях на месте главного слова не может находиться существительное с числовым значением.

Длительные конструкции преобразуются в три вида отношений «advmod», «obl», «nmod». Все три СинтO являются аналогом многих отношений СинTagРус'а, от которых конструкции с длительным СинтO ничем не отличались бы без помет второго уровня (см. «EX. 3.», «EX. 4.», «EX. 5.»). Это является следствием того, что под определение данных отношений подходят множества СинтO из СинTagРус'а: «obl» – для любых неядерных именных актантов, которые в СинTagРус'е могут быть связаны 40 видами СинтO (см. «EX. 8.», «EX. 9.», «EX. 10.», «EX. 11.», «EX. 12.»); «advmod» - для любых наречий, характеризующих предикат или определение, которые могут быть участниками 23 видов СинтO («EX. 6.», «EX. 7.»); «nmod» - для существительных характеризующих другие существительные, подобные слова могут участвовать в 37 отношениях СинTagРус'а («EX. 13.», «EX. 14.», «EX. 15.»). Кроме длительных в отношении «obl» преобразуется кратко-длительное («EX. 16.»), а также

дистанционное СинтО («EX. 17.»), которые теперь также имеют свою помету в разметке UD.

Таблица 4.7. Примеры 2

	СинТАрРус'	UD	Пример
EX. 3.	длительн	obl:dur	Он все время [Y] держит [X] внутреннюю агрессию в состоянии температуры кипения
EX. 4.	длительн	advmod:dur	При этом человек может удержать [X] это изображение в своем представлении столько [Y] времени сколько ему нужно.
EX. 5.	длительн	nmod:dur	Я в театре [X] всю жизнь [Y].
EX. 6.	обст	advmod	Поэтому [Y]. в пятницу к парому выстраивается [X] гигантская многочасовая очередь по 40 50 машин
EX. 7.	1-компл	advmod	Американцы, которые привыкли считать деньги, поступают [X] именно так [Y], отсекая 90 процентов трат на флюорографию.
EX. 8.	обст	obl	Тот же концерн BMW реализует [X] в России [Y] несколько тысяч машин в год, тогда как на рынке Северной Америки счет продаж идет на сотни тысяч.
EX. 9.	1-компл (1) 2-компл(2) 4-компл(3)	obl	Тянет [X3] за шарф [Y3] - он разматывается, хватает [X2] за рукав [Y2], и из кармана [Y1] высыпаются [X1] гвозди.
EX. 10	присвяз	obl	При деформации пластины посадка самолета, естественно, будет [X] невозможна [Y], ведь взлетно-посадочная полоса должна быть ровной.
EX. 11	огранич	obl	Для того чтобы свет распространялся на такие расстояния, полагает Чернобров, тоннель должен быть идеально [Y] прямым [X] и с большим сечением, а его стенки иметь хорошую отражающую поверхность.
EX. 12	суб-копр	obl	Вы же сами [Y] понимаете [X], как это мне трудно.
EX. 13	квазиагент	nmod	Хотя пол года назад во время визита [X] французского министра [Y] образования речь об этом шла как о деле практически решенном.
EX. 14	атриб	nmod	Вряд ли стоит напоминать, что история [X] науки [Y] прошлого столетия не оправдала вышеописанного мнения теоретиков прошлого.
EX. 15	1-компл (1) 3-компл(2) Квазиагент (3) 2-компл(4) Соотнос (5)	nmod	Они помогают нам обеспечивать чистоту [X3] эксперимента [Y3]: от проблемы [X5] удаления [X1] посторонних [Y1] из помещения, где проводится экзамен, до вопросов [Y5] транспортировки и доставки [X1, 2, 4] информации [Y1] из регионов [Y2] в Москву [Y4].
EX. 16	кратно- длительн	obl:mdur	Больные выстроившиеся в поисках исцеления в длинные очереди готовы были ждать [X] часами [Y].
EX. 17	дистанц	obl:dist	Примерно каждые сто метров [Y] нам приходилось останавливаться [X], выходить из машины и смотреть, можно ли проехать дальше.

Изъяснительное СинтО, конвертируемое в parataxis:relcl (см. «EX. 18.», «EX. 19.»), теперь также выделяется среди 6 других отношений, конвертируемых в «parataxis» (см. EX. 20. EX. 21.).

Таблица 4.8. Примеры 2

	СинTagРус'	UD	Пример
EX. 18.	Изъяснительное	parataxis:relcl	В устройстве отсутствуют [X] подвижные части, что позволяет [Y] удешевить производство.
EX. 19.	Изъяснительное	parataxis:relcl	Появлялась она всегда со стороны поля и затем кружилась в пространстве между горелой сосной и той, под которой ребята видели [X] однажды лисицу, отчего и называли [Y] ее "лисичкой".
EX. 20.	разъяснит	parataxis	Математики могут с помощью формул описать все [X] - работу [Y] механизмов, движение космических тел, жизнь человека.
EX. 21.	ввод	parataxis	"Установление "высоты" такого энергетического барьера - это задача [X] науки, одна из最难нейших и важнейших, - считает [Y] академик Григорян.

Отношение «case:compar» теперь оформляет конструкции со сравнительными союзами (см. «EX. 22.», «»).

Таблица 4.9. Примеры 3

	СинTagРус'	UD	Пример
EX. 22.	case	case:compar	Взлетевшая далеко ракета не поднялась над гребнем, только осветила край низкого неба, словно [Y] из-за туч [X].
EX. 23	case	case:compar	А по ту сторону рампы пульсировал и взыхал, как [Y] одно тысячеликое мохнатое существо [X], пятиярусный зрительный зал, который пришел смотреть именно на это чудо.

Вместе с морфологическими пометами «*cstr*» для сравнительной и «*sup*» для превосходной степени прилагательных и наречий новое отношение «case:compar» покрывает большую часть сравнительных конструкций. Невыделенными остаются сравнительные конструкции со следующими языковыми средствами²:

- конструкции со сравнительными предлогами («типа», «наподобие», «вроде», «в отличие от»);

² Летучий А.Б. Сравнительные конструкции // Материалы к корпусной грамматике русского языка. Выпуск II. Синтаксические конструкции и грамматические категории. Под редакцией: В. А. Плунгян, Н. М. Стойнова .. - М.: Издательство Нестор-История, 2017. - С. 134-141.

- конструкции с творительным падежом («быть ключом», «упал камнем в воду»);
- конструкции с глаголами («ходить», «различаться» и т.д.) и прилагательными («схожий»);
- конструкции с сочинительными союзами «и» и «а» («Корабль огромный, и груз немаленький», «Один гусь серый, а другой гусь белый»).

5 Заключение

В своей работе 1999 года И.А. Мельчук пишет по поводу универсальности синтаксических отношений, что «поскольку любое конкретное ПСО представляет собой соответствие между определенными формальными конструкциями данного языка и определенным (присущим данному языку) комплексом смысловых отношений, оно в принципе сугубо «национально», неповторимо» [23.]. Но несмотря на то, что в целом вопрос о реальности создания универсальной грамматики для задач обработки естественных языков едва ли можно считать закрытым, невозможно отрицать наличие относительно успешных попыток применения универсальных разметок. Пример UD наглядно показывает, что многие языковые явления действительно можно усреднить до ограниченного набора категорий, признаков и связей, и тем не менее применимость данного набора порой носит условный характер из-за разнообразия языковых строев и явлений. В данном случае лингвоспецифичные пометы являются подходящим решением, т. к. с одной стороны они позволяют уточнить универсальную разметку в тех случаях, когда уравнивание привело к искажению реального строения языка, а также просто дополнить разметку более детальной информацией о языке, если считается, что без нее упускаются лингвоспецифичные особенности. С другой стороны, использование лингвоспецифичных помет в парсерах является optionalным и

настраиваемым. Таким образом разметку UD можно относительно легко подстраивать под нужды конкретной задачи.

При выполнении данной работы не было цели сделать из UD СинТагРуси восстановить все его отношения, потерянные при конвертации. UD и СинТагРус' это разные проекты, создатели, которых преследовали разные цели. Однако в русском корпусе UD потребовались лингвоспецифичные пометы, чтобы используемая разметка могла более точно описывать явления, происходящие в языке и не уравнивать отличающиеся аспекты. Поэтому разметка СинТагРус'a, изначально построенная для русского языка, существенно облегчает данную задачу, ведь многие СинтО данного корпуса сами по себе описывают особенности русского языка. Тем не менее, существует еще много возможностей для развития в данном направлении. Например, в будущих обновлениях возможно добавить существующую морфологическую помету в UD «Reflex=Yes» для выделения категории возвратности. В украинском корпусе UD используется помета «acl:adv» для наречий, которые по выполняемой в конструкциях роли похожи на прилагательные. Данное отношение может быть применимо и в русском языке. Например, «лестница наверх», «инициатива снизу».

Также параллельно с данной работой в данный момент реализуются следующие пометы второго уровня: обособленно-аппозитивное, номинативно-апозитивное, дательно-субъектное.

6 Список литературы

1. de Marneffe, Marie-Catherine and Manning, Christopher D. "Stanford typed dependencies manual." [Электронный ресурс]. 2008. – Режим доступа: http://nlp.stanford.edu/software/dependencies_manual.pdf. A Universal Part-of-Speech Tagset (google_pos)

2. de Marneffe, Marie-Catherine and Manning. The Stanford typed dependencies representation// CrossParser '08: Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation August – Stroudsburg: Association for Computational Linguistics, 2008. – C. 1–8
3. Joakim Nivre. Towards a Universal Grammar for Natural Language Processing // CICLing 2015, Part I, LNCS 9041. - Berlin: Springer International Publishing Switzerland 2, 2015. - C. 3–16.
4. Joakim Nivre, Chiao Ting Fang. Universal Dependency Evaluation // Proceedings of the NoDaLiDa Workshop on Universal Dependencies, UDW@NoDaLiDa. - Gothenburg: Association for Computational Linguistics, 2017. - C. 86 - 95.
5. K. Droganova, D. Zeman. Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies. - Praha: ÚFAL MFF UK, 2016. - 47 c.
6. Lyashevskaya Olga, Droganova Kira, Zeman Daniel, Alexeeva Maria, Gavrilova Tatiana, Mustafina Nina, Shakurova Elena. Universal Dependencies for Russian: A New Syntactic Dependencies Tagset // SSRN Electronic Journal. 2016. 10.2139/ssrn.2859998. – Режим доступа:
https://www.researchgate.net/publication/323980060_Universal_Dependencies_for_Russian_A_New_Syntactic_Dependencies_Tagset
7. Matthew S. Dryer Order of Numeral and Noun, Dryer, Matthew S. & Haspelmath, Martin (eds.) [Электронный ресурс]. – Электрон. дан. – Leipzig: Max Planck Institute for Evolutionary Anthropology. – Режим доступа: <http://wals.info/chapter/89>, Accessed on 2020-05-09
8. Osborne, Timothy, Gerdes, Kim. The status of function words in dependency grammar: A critique of Universal Dependencies (UD) // Glossa: a journal of general linguistics. 2019. 4(1):17. – Режим доступа:
https://www.researchgate.net/publication/330740348_The_status_of_fun

- ction_words_in_dependency_grammar_A_critique_of_Universal_Dependencies_UD (дата обращения: 10.05.2020)
9. Petrov, Slav, Das, Dipanjan, McDonald A Universal Part-of-Speech Tagset. // ResearchGate URL:
https://www.researchgate.net/publication/51887367_A_Universal_Part-of-Speech_Tagset (дата обращения: 10.05.2020)
 10. R. Bivon. Element order. (Studies in the Modern Russian Language 7.) London: Cambridge University Press, 1971. Pp. 86.
 11. Sabine Buchholz, Erwin Marsi. CoNLL-X shared task on multilingual dependency parsing // Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X). - New York City: Association for Computational Linguistics, 2006. - С. 149–164.
 12. Zeman, Daniel. Reusable Tagset Conversion Using Tagset Drivers. [Электронный ресурс]. 2008. – Режим доступа:
https://www.researchgate.net/publication/220746946_Reusable_Tagset_Conversion_Using_Tagset_Drivers
 13. Голуб И.Б. Стилистика русского языка. – М.: Айрис-пресс, 2002.
 14. 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) (рус.) // Сборник «Национальный корпус русского языка: 10 лет проекту». — М.: Труды Института русского языка им. В.В. Виноградова, 2015. — Вып. 6. — С. 272—299.
 15. Е.С. Иншакова, Л.Л. Иомдин, Л.Г. Митюшин, В.Г. Сизов, Т.И. Фролова, Л.Л. Цинман. СинТагРус' сегодня. // Труды Института русского языка им. В.В. Виноградова, т. 21, М., 2019. С 14-41.
 16. Иомдин, Л.Л. Автоматическая обработка текста на естественном языке: модель согласования / Л.Л. Иомдин. М.: Наука, 1990.
 17. Лаврик Эльвира Петровна Особенности согласования в аппроксимативных конструкциях // Известия ВГПУ. 2008. №10. –

Режим доступа: <https://cyberleninka.ru/article/n/osobennosti-soglasovaniya-v-approksimativnyh-konstruktsiyah> (дата обращения: 10.05.2020).

18. Летучий А.Б. Сравнительные конструкции // Материалы к корпусной грамматике русского языка. Выпуск II. Синтаксические конструкции и грамматические категории. Под редакцией: В. А. Плунгян, Н. М. Стойнова . - М.: Издательство Нестор-История, 2017. - С. 132-205.
19. Литвинова Т. Н. Системная характеристика сложноподчинённого изъяснительного предложения // Научные ведомости БелГУ. Серия: Гуманитарные науки. 2014. №13 (184). – Режим доступа:: <https://cyberleninka.ru/article/n/sistemnaya-harakteristika-slozhnopodchinyonnogo-izyasnitelnogo-predlozheniya> (дата обращения: 10.05.2020).
20. Матвеева Т.В. Функциональные стили в аспекте текстовых категорий. Синхронно-сопоставительный очерк. – Свердловск: Издво Уральского ун-та, 1990.
21. Мельчук И.А. «Русский язык в модели «Смысл Текст». Москва – Вена: Wiener Slavistischer Almanach, 1995. – 714 с.
22. Мельчук И. А. Поверхностный синтаксис русских числовых выражений. Wiener slawistischer Almanach Wiener slawistischer Almanach: Sonderband. - 16 изд. - Wien: Institut für Slavistik der Universität Wien, 1985. - 509 с.
23. Мельчук И.А. Опыт теории лингвистических моделей Смысл ↔ текст : Семантика, синтаксис. - XXII изд. - М.: Издательство: Яз. рус. культуры, 1999. - 345 с.
24. Официальный сайт проекта Universal Dependencies [Электронный ресурс]. – Режим доступа: <https://universaldependencies.org/> (дата обращения: 10.05.2020).

25. Официальный сайт проекта НКРЯ (синтаксический корпус) [Электронный ресурс]. – Режим доступа:
<http://ruscorpora.ru/new/instruction-syntax.html>
26. Парамонов Д.А., Феномен грамматического выражения модальных значений в современном русском языке в свете экспрессивности: дис. ...д-р филол. наук. – М., 2010. – 722 с .
27. Скоринкин Даниил Андреевич. Электронное представление текста с помощью стандарта разметки tei // Вестник Московского университета. Серия 9. Филология. 2016. №5. – Режим доступа:
<https://cyberleninka.ru/article/n/elektronnoe-predstavlenie-teksta-s-pomoschyu-standarta-razmetki-tei> (дата обращения: 10.05.2020).
28. Чайковская Н. Н. Способы выражения изъяснительных отношений в современном русском языке. Диссертации по гуманитарным наукам (1989) // [Электронный ресурс]. – Режим доступа: филол. наук. – М., 2010. – 722 с. <http://cheloveknauka.com/sposoby-vyrazheniya-izyasnitelnyh-otnosheniy-v-sovremennom-russkom-yazyke> (дата обращения: 10.05.2020).
29. Шведова Н.Ю. Русская грамматика. - Т.2. изд. - М.: Издательство «Наука», 1980. - 717 с.
30. Шведова Н.Ю. Русская грамматика. - Т.2. изд. - М.: Издательство «Наука», 1980. - 792 с.