Создание корпуса русских текстов с разметкой систем синтаксических групп

Д.В. Демидов, И.В. Евченко

***Аннотация* — В работе проанализированы способы представления синтаксической структуры предложения. Выделены недостатки формализмов и основные проблемы построения структуры предложения. Описана технология формирования синтаксически размеченного корпуса русских текстов, в котором синтаксическая структура предложений представляет собой размеченную систему синтаксических групп (ССГ) Гладкого А.В. В качестве исходного материала для корпуса служит СинТагРус. В статье описан способ применения формата CoNLL-U для представления ССГ и примеры правил выделения синтаксических групп на основе деревьев зависимостей (ДЗ). Представлены программные средства трансформации ДЗ в ССГ и средства визуализации ССГ. Полученные результаты делают возможным создание синтаксического анализатора, строящего ССГ, методами машинного обучения, что не исключает применимость и традиционного подхода.**

***Ключевые слова* — Системы синтаксических групп, деревья зависимостей, синтаксически размеченный корпус текстов, визуализация систем синтаксических групп.**

I. Введение

Широкое распространение методов машинного обучения в современной компьютерной лингвистике порождает потребность в больших объемах аннотированного текстового материала. Для русского языка разрабатываются морфологически и синтаксически размеченные корпусы [2], [3], [4].

Известно, что языковой корпус является одновременно инструментом для анализа и разработки лингвистических компонент, для которых используют синтаксическую разметку (синтаксис составляющих, синтаксис зависимостей). Анализ существующих подходов к построению корпусов показал, что для грамматики составляющих не учитывается тот факт, что в языках со свободным порядком слов составляющая может разрываться, а для грамматики зависимостей существует ряд проблем, касающихся сочинительных, придаточных предложений. Чтобы исправить недостатки последнего подхода Гладкий А.В. в работе [1] предложил системы синтаксических групп (ССГ).

ССГ занимают промежуточное место между деревьями зависимостей (ДЗ) и деревьями непосредственных составляющих (ДНС), избегая ряда недостатков обоих формализмов [1], [6].

Однако, для такого формализма как системы синтаксических групп Гладкого А.В., по всей видимости, не существует ни корпуса ССГ, ни программной реализации синтаксического анализатора.

Сложность аппарата ССГ создаёт значительные препятствия для алгоритмизации синтаксического анализа, поэтому задача предварительного формирования корпуса текстов представляется более актуальной — доступность подобного корпуса открыло бы возможность применения методов машинного обучения для построения синтаксических анализаторов в том числе.

В данной статье рассматривается технология формирования корпуса русских текстов с разметкой систем синтаксических групп на основе корпуса текстов с синтаксической разметкой в виде ДЗ.

В качестве основного метода построения систем синтаксических групп используется аппарат, разработанный А. В. Гладким. Размеченная система синтаксических групп — граф, узлы которого представляют собой синтаксические группы (СГ) — подмножества множества всех слов предложения, а дуги соответствуют отношению подчинения между узлами [1]. Будем размечать СГ кодами подкритериев, в соответствии с которыми они выделяются, а дуги — синтаксическими отношениями, принятыми в СинТагРусе.

II. Представление систем синтаксических групп

Формат CoNLL-U разработан для представления размеченных ДЗ в табличном виде с морфологической информацией для слов. В этом формате можно представить одно или несколько предложений, где каждому слову предложения отводится одна строка. Колонки таблицы табулированы. ДЗ для предложения «*Он наконец нашел и кремень, и кусок напильника, и шнур в патронной гильзе, вымокший в снегу*» приведено в таблице 1, в которой для наглядности опущены столбцы FEATS, DEPS, MISC.

Из соображений совместимости в этом же формате удобно представлять и ССГ, но с введением ряда договорённостей:

1) Символом подчёркивания в поле DEPREL будем обозначать отношение «часть-целое» между словом (или СГ) и непосредственно включающей его СГ.
2) Для СГ поля FORM и LEMMA будут содержать знаки подчёркивания, в отличие от обычных слов.

Статья получена 26 января 2024.
Дмитрий Витальевич Демидов, Московский инженерно-физический институт (национальный исследовательский ядерный университет), (e-mail: dvdemidov@mephi.ru).
Игорь Владимирович Евченко, Московский инженерно-физический институт (национальный исследовательский ядерный университет), (e-mail: t.foreli@ya.ru).

Таблица 1. Примера дерева зависимостей в формате CoNLL-U

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	HEAD	DEPREL
1	Он	ОН	_	_	3	предик
2	наконец	НАКОНЕЦ	_	_	3	обст
3	нашел	НАХОДИТЬ	_	_	0	_
4	и	И	_	_	6	соотнос
5	кремень,	КРЕМЕНЬ	_	_	3	1-компл
6	и	И	_	_	5	сочин
7	кусок	КУСОК	_	_	6	соч-союзн
8	напильника,	НАПИЛЬНИК	_	_	7	квазиагент
9	и	И	_	_	7	сочин
10	шнур	ШНУР	_	_	9	соч-союзн
11	в	В	_	_	10	атриб
…	…	…	…	…	…	…

3) Поле UPOSTAG содержит саму СГ, что интерпретируется специальными средствами визуализации.

4) Поле XPOSTAG содержит код подкритерия, по которому выделена СГ, а для всего предложения здесь задействуется символ '*'.

5) Поля HEAD и DEPREL сохраняют свой смысл.

Пример ССГ приведён в таблице 2.

Заметим, что существующие средства визуализации ДЗ в формате CoNLL-U будут генерировать правильно построенные деревья и для ССГ, но не будут интерпретировать отношение «часть-целое» должным образом. Поэтому для визуализации ССГ разработаны средства визуализации, которые способны отображать как ДЗ, так и ССГ, согласно подходу в [6].

Таблица 2. Фрагмент ССГ в формате CoNLL-U

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	HEAD	DEPREL
1	_	_	<всё предложение>	*	0	_
2	Он	ОН	NOUN	S	4	предик
3	наконец	НАКОНЕЦ	ADV	ADV	4	обст
4	нашел	НАХОДИТЬ	VERB	V	1	_
5	и	И	CCONJ	CONJ	6	_
6	_	_	и…и…и	АЗ	4	1-компл
7	кремень,	КРЕМЕНЬ	NOUN	S	6	соч-союзн
8	и	И	CCONJ	CONJ	6	_
9	кусок	КУСОК	NOUN	S	6	соч-союзн
10	напильника,	НАПИЛЬНИК	NOUN	S	9	квазиагент
…	…	…	…	…	…	…

Причём размеченные синтаксические группы отображаются в явном виде, а отношения «часть-целое» графически оформляются пунктирной стрелкой, отличаясь от отношений зависимости, как показано на рисунке 1.

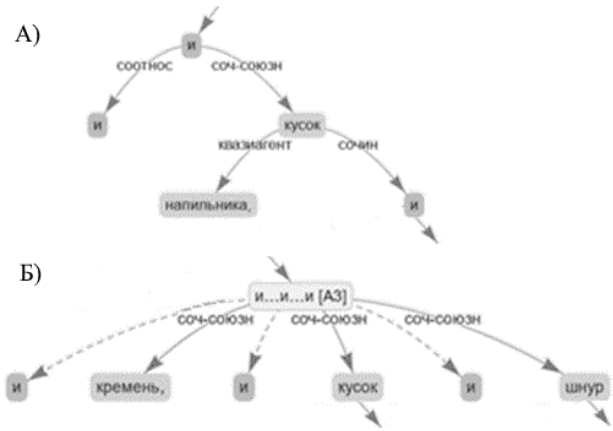

Рисунок 1. Пример синтаксической структуры, А) ДЗ, Б) ССГ

III. Технология построения ССГ

Для автоматизированной разметки корпуса русских текстов используется синтаксически размеченный корпус русских текстов (СинТагРус), разработанный лабораторией компьютерной лингвистики ИППИ РАН. [5]

Для выделения различных СГ используются критерии Гладкого А.В. В ходе работы для большинства из них предложен набор процедур и правил трансформации деревьев зависимостей.

Применение этих процедур и правил выполняется итеративно, начиная с исходного дерева зависимостей, — на очередной итерации то или иное правило трансформирует текущее частично построенное представление.

Процесс завершается, когда обработаны все критерии и ни одно из правил не может быть применено.

Одно и то же правило может применяться многократно. Обработка правил совершается последовательно с учётом приоритета подкритериев. В результате цепочки трансформаций строится корректная синтаксическая структура, как показано на рисунке 2.

Приоритизация подкритериев осуществлялась на основе результатов статистической обработки синтаксических структур предложений в примерах книги Гладкого А.В.

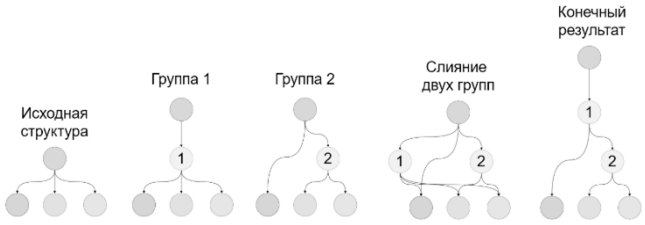

Рисунок 2. Пример трансформации синтаксической структуры

По итогам анализа был построен ориентированный граф зависимостей подкритериев (рисунок 3), согласно которому, например, подкритерии Г5, Г6, Г7, Г13, Г15 не зависят от других подкритериев, т. к. в эти вершины не входят стрелки. А подкритерии А2, А8, Б3 зависят от других подкритериев, и от них не зависят другие подкритерии, т. к. из соответствующих вершин не выходят стрелки к другим вершинам.

При исследовании графа на частичную упорядоченность было замечено, что в графе присутствуют циклы, например, Б2→Б1→Г9→Г14→Б2. Это означает, что в общем случае недостаточно единожды применить подкритерии циклической группы, т. к. в ходе трансформаций синтаксической структуры могут создаться условия повторного применения этих подкритериев.

Разумным условием прекращения применения подкри-

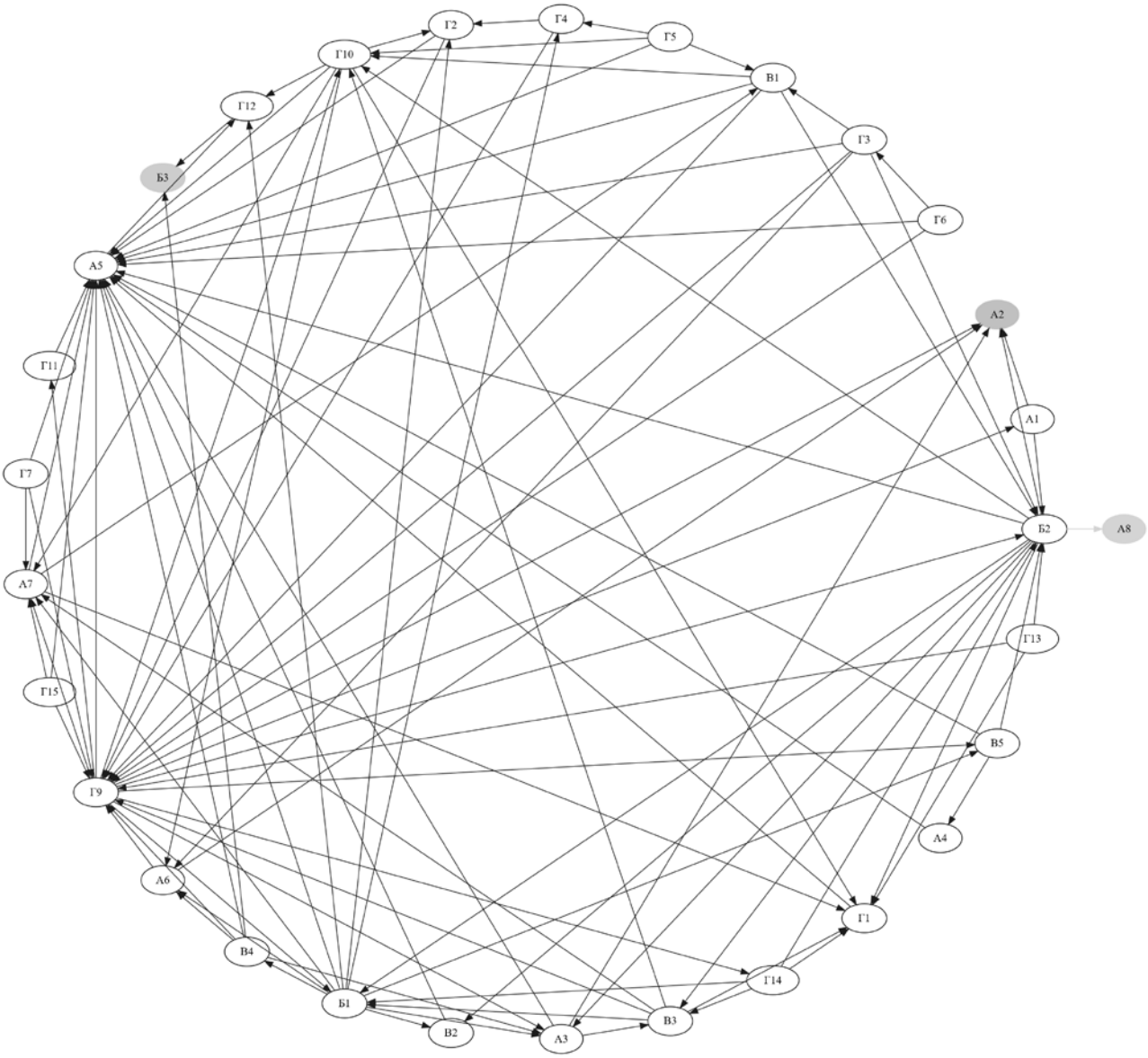

Рисунок 3. Ориентированный граф зависимостей подкритериев Гладкого А.В.

териев циклической группы является отсутствие изменений в синтаксической структуре после очередного прохода по всем подкритериям циклической группы.

Некоторые подкритерии представляют особенную сложность в автоматизации. Например, подкритерий Г1, который включает в себя конструкции вроде «высокого роста», «с живыми глазами», «в отличном расположении духа» и т. п., в которых главный член не может употребляться без зависимого.

Отдельное место в корпусе ССГ занимают вручную размеченные примеры ССГ из [1].

IV. Правила выделения синтаксических групп

Ради наличия корня синтаксического дерева для каждого предложения формируется одна общая СГ, которая является тривиальной группой, соответствующей всему предложению. Остальные СГ формируются в соответствии с критериями А, Б, В, Г из [1], представленными в виде набора подкритериев.

В качестве примеров рассмотрим алгоритмы и общие схемы для некоторых подкритериев.

В соответствии с критерием А существует подкритерий А1, выделяющий группу сложных и составных предлогов («в течение», «в продолжение», «в дополнение к» и др.). Если в предложении есть сложные и составные предлоги, то в предложении они идут последовательно. Поэтому в структуре дерева такие токены всегда находятся рядом друг с другом и связаны определённой зависимостью. Определив их местоположение, можно построить СГ и расположить её в дереве над найденными токенами.

Шаги алгоритма:

1) Найти предложную связь.
2) Образовать синтаксическую группу из частей предлога.
3) Входящая связь с главного слова переключается на образованную группу.
4) Связь между главным и зависимым словом удаляется.
5) Главное и зависимое слово связываются с созданной группой отношениями «часть-целое».
6) Все исходящие из зависимого слова связи передаются созданной группе.

Визуализация фрагмента дерева (А) после отработки описанным алгоритмом показана на рисунке 2 (Б).

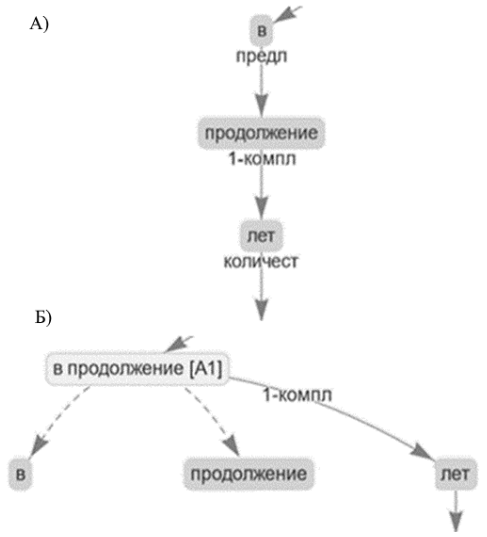

Рисунок 2. Пример синтаксической структуры, А) ДЗ, Б) ССГ

Следующими можно отметить подкритерии Б1 и В2, работающие вместе. Словосочетания, входящие в состав простого предложения или одной из компонент сложного и вводимые союзами, выделяются в группу по критерию В2. «Внешние» СГ, включающие союз, выделяются в СГ по пункту Б1, как неподчинительные контексты

Идея алгоритма в обнаружении сравнительной связи между токенами и создании группы, убедившись в том, что в структуре дерева такие токены выполняют роль сравнения и являются одной из компонент сложного предложения. Шаги алгоритма:

1) Найти сравнительную связь.
2) Проверить наличие компонентов сравнительной связи и вложенные компоненты сложного предложения.
3) Образовать синтаксическую группу из частей сравнения и компонентов сложного предложения.
4) Входящая связь с главного слова переключается на образованную группу.
5) Связь между главным и зависимым словом удаляется.
6) Главное и зависимое слово связываются с созданной группой отношениями «часть-целое».
7) Все исходящие из зависимого слова связи передаются созданной группе.

Визуализация примера после работы описанного алгоритма представлена на рисунке 3.

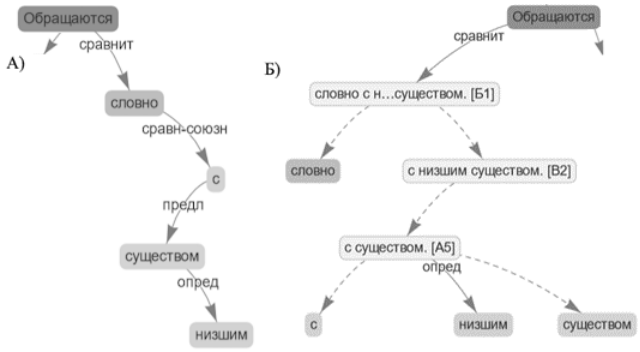

Рисунок 3. Пример синтаксической структуры до и после обработки, А) ДЗ, Б) ССГ

Г критерий содержит намного больше подкритериев, среди которых рассмотрим Г1. Он выделяет конструкции вроде «высокого роста», «с живыми глазами» и т. п., в которых главный член не может употребляться без зависимого.

Алгоритм создания группы отличается от предыдущих, так как для этого типа необходим достаточно большой список словосочетаний, которые подходят под особенные конструкции, выделяемые критерием Г.

Шаги алгоритма:

1) Согласно списку словосочетаний, найти такие словосочетания в предложении.
2) Образовать группу.
3) Входящая связь для главного слова меняется и становится входящей связью для образованной группы.
4) Связь между главным и зависимым словом удаляется.
5) Главное и зависимое слово связываются с созданной группой отношениями «часть-целое».
6) Все исходящие из зависимого слова связи передаются созданной группе.

Пример предложения, в котором удалось выделить группу по этому подкритерию, показан на рисунке 4.

Стоит отметить, что синтаксическая структура предложения в итоге может содержать в себе одновременно множество разных групп.

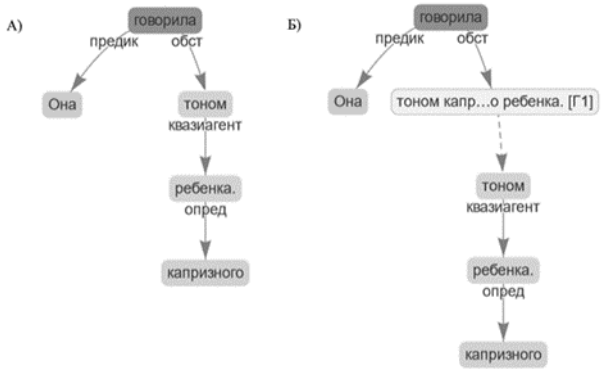

Рисунок 4. Пример синтаксической структуры предложения «Она говорила тоном капризного ребенка», А) ДЗ, б) ССГ

V. Анализ предварительных результатов

На сколько сложнее или проще ССГ по сравнению с деревьями зависимостей (ДЗ) и деревьями непосредственных составляющих (ДНС)?

Для ответа на этот вопрос необходимо ввести некоторый количественный показатель сложности, который можно было бы рассчитать для всех трёх представлений.

Определим индекс сложности синтаксической структуры (ИССС) как отношение числа элементов синтаксической структуры к числу слов в предложении. Для ДЗ в числителе будет количество токенов и зависимостей; для ДНС — количество составляющих, включая слова; для ССГ — количество токенов, СГ и зависимостей.

Для 266 предложений из 2175 слов были построены ДЗ с общим числом зависимостей 1974 и ССГ с 819 СГ и 1579 зависимостями. Число зависимостей в ССГ оказалось на 20% меньше числа зависимостей в ДЗ за счёт замещения ряда синтаксических отношений синтаксическими группами. В частности, полностью исчезли сентенциальные, сочинительные, союзные, аналитические, подчинительные и предложные отношения; частично исчезли количественные, комплетивные, обстоятельственные, определительные, предикативные отношения. С другой стороны, общее количество СГ и зависимостей в ССГ составило 2398, что превышает число зависимостей в ДЗ на 21%.

ИССС для рассмотренных ССГ составил в среднем 2.1, для ДЗ — 1.9 (при этом для СинТагРуса этот показатель равен 1.92). Для рассмотренных предложений ДНС не строились, но удалось рассчитать ИССС для корпуса текстов американского английского языка Penn Treebank [8], содержащего 260 тысяч предложений с разметкой непосредственных составляющих. Он составил 2.58.

Полученные значения ИССС позволяют утверждать, что аппарат ССГ лишь незначительно «утяжеляет» синтаксическую структуру по сравнению с ДЗ, устраняя при этом ряд недостатков ДЗ и привнося полезные свойства ДНС.

VI. Заключение

Предложенная технология позволяет выделять синтаксические группы в предложении на основе дерева зависимостей. Как следствие, создание синтаксического корпуса русских текстов с разметкой систем синтаксических групп может выполняться на основе существующих синтаксически размеченных корпусов, таких как СинТагРус. Наибольшая полезность технологии видится в том, что открывается возможность для создания синтаксических анализаторов, строящих ССГ, и автоматической оценки их качества работы.

Работа над корпусом ещё не завершена. На настоящий момент подготовлены алгоритмы для 17 подкритериев, что составляет примерно половину от планируемого объёма. В будущем планируется усовершенствовать алгоритмы и правила выделения синтаксических групп до достижения полноты, а далее работать над точностью.

Далее, созданный корпус русских текстов с разметкой систем синтаксических групп планируется использовать для машинного обучения синтаксического анализатора. Авторы надеются, что корпус послужит материалом для широкого спектра исследований как фундаментального, так и прикладного характера.

Библиография

[1] Гладкий А.В. Синтаксические структуры естественного языка. Изд. 3-е, стереотип. М.: ЛЕНАНД. 2018. — 152 с.

[2] Droganova K., Zeman D. Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies // Technical report. — Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, 2016.

[3] Иншакова Е., Иомдин Л., Митюшин Л., Сизов В., Фролова Т., Цинман Л. СинТагРус сегодня. Proceedings of the V.V. Vinogradov Russian Language Institute. 2019.

[4] Власова Н.А., Трофимов И.В., Сердюк Ю.П., Сулейманова Е.А., Воздвиженский И.Н. PaRuS - синтаксически аннотированный корпус русского языка // Программные системы: теория и приложения. 2019. №4 (43).

[5] Kibrik, Andrej A., Dobrov, Grigory B., & Korotaev, Nikolay A. Modelling natural communication and a multichannel resource: The deceleration effect // V. Solovyev, N. Loukachevitch, O. Lyashevskaya (Eds.) Proceedings of the Linguistic Forum 2020: Language and Artificial Intelligence. Moscow, Russia, November 12–14, 2020. 2021.

[6] Демидов Д.В. Представление синтаксических структур с сочинительными конструкциями // Искусственный интеллект и принятие решений. №2. 2022. С. 36–50.

[7] Коротаев Н. А. Синтаксические группы А. В. Гладкого: анализ конструкций с сочинением // Вестник РГГУ. Серия: Литературоведение. Языкознание. Культурология. 2013. №8 (109).

[8] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: the penn treebank. Comput. Linguist. 1993, PP. 313–330.

Creating a corpus of Russian texts with markup of syntactic group systems

Dmitriy V. Demidov, Igor V. Evchenko

***Abstract* — The paper analyzes the ways of representing syntactic structure of a sentence. The disadvantages of formalisms and major problems in the construction of sentence' syntactic structure are highlighted. The technology of forming a syntactically tagged corpus of Russian texts is described, in which the syntactic structure of sentences is represented by the tagged system of syntactic groups (SSG) of Gladky A.V. SynTagRus serves as a source material for the corpus. The paper describes a method of using the CoNLL-U format for representing SSGs and examples of rules for selecting syntactic groups based on dependency trees. Software tool for transformation of dependency trees into SSGs and tool for visualizing SSGs are presented. The results obtained make it possible to create a syntactic parser that builds SSGs using machine learning methods not excluding the applicability of the traditional approach.**

***Keywords* — Syntactic group systems, dependency trees, syntactically tagged corpus of texts, visualization of systems of syntactic groups.**

REFERENCES

[1] Gladky A.V. Syntactic structures of natural language. Ed. 3rd, stereotype. M.: LENNAND. 2018. 152 p.

[2] Droganova K., Zeman D. Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies // Technical report. — Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, 2016.

[3] Inshakova E., Iomdin L., Mityushin L., Sizov V., Frolova T., Tsinman L. SinTagRus today. Proceedings of the V.V. Vinogradov Russian Language Institute. 2019.

[4] Vlasova N.A., Trofimov I.V., Serdyuk Yu.P., Suleymanova E.A., Vozdvizhensky I.N. PaRuS - syntactically annotated corpus of the Russian language // Software systems: theory and applications. 2019. No. 4 (43).

[5] Kibrik, Andrej A., Dobrov, Grigory B., & Korotaev, Nikolay A. Modeling natural communication and a multichannel resource: The deceleration effect // V. Solovyev, N. Loukachevitch, O. Lyashevskaya (Eds.) Proceedings of the Linguistic Forum 2020: Language and Artificial Intelligence. Moscow, Russia, November 12-14, 2020. 2021.

[6] Demidov D.V. Representation of syntactic structures with coordinating constructions // Artificial intelligence and decision making. No. 2. 2022. pp. 36-50.

[7] Korotaev N. A. Syntactic groups of A. V. Gladky: analysis of constructions with an essay // Bulletin of the Russian State University for the Humanities. Series: Literary Studies. Linguistics. Culturology. 2013. No. 8 (109).

[8] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: the penn treebank. Comput. Linguist. 1993, PP. 313–330.