

Институт интеллектуальных кибернетических систем



КАФЕДРА КИБЕРНЕТИКИ

## Задание на УИР

Студенту Б20-514 Костылев Вадим Дмитриевич  
гр. \_\_\_\_\_  
(группа) (фио)

### ТЕМА УИР

Исследование возможности автоматизации выделения синтаксических групп по  
термово-контекстному критерию

### ЗАДАНИЕ

№ п/п	Содержание работы	Форма отчетности	Срок исполнения	Отметка о выполнении Дата, подпись
1.	<b>Аналитическая часть</b>			
1.1.	Анализ подхода А.В. Гладкого к выделению синтаксических групп с помощью термово-контекстного критерия	Раздел ПЗ		
1.2.	Анализ возможности автоматического выделения СГ по термово-контекстному критерию из предложения	Раздел ПЗ		
1.3.	Анализ представлений размеченных синтаксических структур с учётом омонимии	Раздел ПЗ		
1.4.	Оформление расширенного содержания пояснительной записки (РСПЗ)	Текст РСПЗ		
2.	<b>Теоретическая часть</b>			
2.1.	Разработка алгоритма автоматического выделения СГ по термово-контекстному критерию из дерева зависимостей	Раздел ПЗ		
2.2.	Разработка правил трансформации деревьев зависимостей при выделении синтаксических групп по термово-контекстному критерию	Диаграммы, псевдокод, описание		
3.	<b>Инженерная часть</b>			
3.1.	Проектирование и реализация программных модулей выделения синтаксических групп по термово-контекстному критерию из дерева зависимости	Описание архитектуры,		

		исходные тексты		
3.2.	Построение производных синтаксических структур по деревьям зависимости Синтагруса	Тестовые примеры		
4.	<b>Технологическая и практическая часть</b>			
4.1.	Интеграция в Программу преобразования деревьев зависимостей в гибридные синтаксические структуры с синтаксическими группами	Техническое описание		
4.2.	Построение систем синтаксических групп по деревьям зависимости Синтагруса	Описание экспериментов		
	<i>Оформление пояснительной записки (ПЗ) и иллюстративного материала для доклада.</i>	Текст ПЗ, презентация		

## ЛИТЕРАТУРА

1.	Гладкий А.В. Синтаксические структуры естественного языка. Изд. 3-е, стереотип. М.: ЛЕНАНД. 2018. 152 с.
2.	Коротаев Н. А. Синтаксические группы А.В. Гладкого: анализ конструкций с сочинением // Вестник РГГУ. Серия «Филологические науки. Языкознание / Московский лингвистический журнал». Том 13. 2013. С. 16-36.
3.	Падучева Е.В. О семантике синтаксиса: материалы к трансформационной грамматике русского языка. Изд. стереотип. М.: Книжный дом «ЛИБРОКОМ». 2019. 296 с.
4.	Мельчук И.А. Автоматический синтаксический анализ. Новосибирск. 1964. 364 с.
5.	Демидов Д.В. Представление синтаксических структур с сочинительными конструкциями // Искусственный интеллект и принятие решений, 2022г. Вып. 2 Стр. 36-50

Дата выдачи задания:

Руководитель

Демидов Д.В.

(ФИО)

« 7 » февраля 2023 г.

Студент

Костылев В.Д.

(ФИО)

## Оглавление

1. Введение.....	4
2. Аналитическая часть.....	5
3. Анализ подхода А.В. Гладкого к выделению синтаксических групп с помощью термово- контекстного критерия.....	5
4. Анализ возможности автоматического выделения СГ по термого-контекстному критерию из предложения.....	6
5. Анализ представлений размеченных синтаксических структур с учётом омонимии.....	6
6. Теоретическая часть.....	7
7. Разработка алгоритма автоматического выделения СГ по термово-контекстному критерию из дерева зависимостей.....	7

## **Введение**

Для разработки автоматизированных систем общения и для других задач, связанных с обработкой текста на естественном языке требуется описать структуру предложения. Для решения данной задачи можно использовать разбиение предложений на синтаксические группы, критерии для чего предложены Гладким Алексеем Всеволодовичем.

Реализовываться будет критерий Г, а целью работы будет модуль для программного средства для анализа уже размеченных предложений.

## 1. Аналитическая часть

### 1.1 Анализ подхода А.В. Гладкого к выделению синтаксических групп с помощью термово-контекстного критерия

Критерий Г (термово-контекстный критерий) - наиболее продуктивный для выделения ССГ критерий, 18 типов которых описаны Гладким.

Критерии:

1. Конструкции вроде *высокого роста, с живыми глазами*, в которых главный член не может употребляться без зависимого.
2. Конструкции вроде *пришел уставший, сидел в шубе*, в которых отсутствует семантическая связь между главным и зависимым словом
3. Конструкции глагол→инфинитив, в случае если у инфинитива есть зависимое, которое в случае, если у словосочетания есть слово, которое может быть истолковано как зависимое от инфинитива, и это приведет к нарушению проективности.
4. В аналогичных условиях выделяются существительное, прилагательное или предикативный перечень→инфинитив.
5. В аналогичных условиях выделяются  $(V) \rightarrow (X) \rightarrow I \rightarrow \dots \rightarrow I$  (V - Глагол, X - существительное, прилагательное, предикативное наречие, I - инфинитив повторяется n раз  $n \geq 1$ ).
6. В аналогичных условиях выделяются  $(V) \rightarrow (X) \rightarrow I \rightarrow \dots \rightarrow I \rightarrow N$  (V - Глагол, X - существительное, прилагательное, предикативное наречие, I - инфинитив повторяется n раз  $n \geq 1$ , N - имя).
7. Сочетание вида  $V \rightarrow N$  или  $(V) \rightarrow (X) \rightarrow I \rightarrow \dots \rightarrow I \rightarrow N$ , выделяется также в том случае, если имеет зависимое словосочетание, которое может подчиняться имени лишь находясь в постпозиции или предпозиции.
8. Во всех случаях когда выделяется в СГ  $V \rightarrow Y$ , Y - что угодно, то выделяется и  $PV \rightarrow Y$ .
9. По критерию Г выделяются в СГ все подчинительные операторные контексты.
10. Частный случай операторного контекста, когда второй член есть асемантический глагол.
11. Контексты со вторыми членами относительно-характеризующими словами и словосочетаниями.
12. Также выделяются количественные кванторы
13. Словосочетание из глагола или зависящего от него слова или словосочетания и имеющее при себе обстоятельство, выделяется в СГ если обстоятельство не может быть без изменения смысла отнесено к глаголу.
14. В сложноподчиненном предложении придаточная часть может не допускать придаточная часть может не допускать соотнесения с каким-либо членом главной части. В таком случае в СГ должна выделяться некоторая группа членов главной части или вся главная часть в целом.
15. Вводные слова и словосочетания.
16. Обращения.
17. Идиоматические словосочетания.
18. Собственные имена, образующие деревья.

## **1.2. Анализ возможности автоматического выделения СГ по термово-контекстному критерию из предложения**

Часть критериев, не получится реализовать. Например критерий 1 (так как нет четких критериев того, как можно выделить словосочетания в которых главный член не употребляются без зависимого, это зависит от смысла сказанного), 13 (точного критерия отсутствия смысла не существует). Также для критерия 17 придется использовать перечисление всех возможных идиом русского языка, а для критерия 8 придется сначала выполнить 3, 4, 5, 6, 7.

## **1.3 Анализ представлений размеченных синтаксических структур с учётом омонимии**

Для решения описанной задачи был предоставлен conllu файл с размеченными предложениями СинТагРус. СинТагРус - фрагмент Национального корпуса, содержащий тексты с синтаксической разметкой. Структура предложений в данном корпусе представляет собой дерево зависимостей, в узлах которого стоят слова предложения, а дуги помечены именами синтаксических отношений. В состав корпуса входят как научно-популярные, общественно-политические и информационные статьи, так и художественная литература.

Как части речи в корпусе выделены:

1. V - галгол
2. S - существительное
3. A - прилагательное
4. ADV - наречие
5. NUM - числительное
6. PR - предлог
7. CONJ - союз
8. PART - частица
9. INTJ - междометие
10. COM - композит
11. NID - иноязычное вкрапление в русский язык

Также в корпусе присутствуют синтаксические отношения, которые разделены на несколько групп: актантные, атрибутивные, сочинительные, служебные.

Форматом файла, который используется для решения задачи является conllu. CoNLL-U - самый популярный и общий используемый формат записи деревьев зависимостей. Файл представляет из себя таблицы, полями которой являются: id(индекс слова), form(сам токен, форма слова), lemma(начальная форма), upos(частеречный тег), xpos(лингвоспецифическая часть речи), feats(морфологическая информация), head(id родителя), deprel(тип зависимости), deps альтернативный подграф в виде пар "вершина-тип зависимости", misc - комментарии и другие пометки.

```

# sent_id = Nachalo.tgt.tgtreg_169
1  _  _  Я прочел стихи Брюсова, которого тогда любил за щедрость звуков.  A0  _  0  _  0:_  _
2  Я  Я  NOUN  S  S  ЕД МУЖ ИМ ОД 3  предик 3:предик  _
3  прочел  ЧИТАТЬ  VERB  V  V  СОВ ИЗЪЯВ ПРОШ ЕД МУЖ 1  _  1:_  _
4  стихи  СТИХИ  NOUN  S  S  МН МУЖ ВИН НЕОД 3  1-компл 3:1-компл  _
5  Брюсова,  БРЮСОВ  NOUN  S  S  ЕД МУЖ РОД ОД 4  квазиагент 4:квазиагент  _
6  которого  КОТОРЫЙ  NOUN  S  S  ЕД МУЖ ВИН ОД 8  1-компл 8:1-компл  _
7  тогда  ТОГДА  ADV  ADV  ADV 8  обст 8:обст  _
8  любил  ЛЮБИТЬ  VERB  V  V  НЕСОВ ИЗЪЯВ ПРОШ ЕД МУЖ 5  релят 5:релят  _
9  за  ЗА  ADP  PR  PR 8  2-компл 8:2-компл  _
10 щедрость  ЩЕДРОСТЬ  NOUN  S  S  ЕД ЖЕН ВИН НЕОД 9  предл 9:предл  _
11 звуков.  ЗВУК  NOUN  S  S  МН МУЖ РОД НЕОД 10  квазиагент 10:квазиагент  _

```

Рис 1. Пример размеченного предложения

## 2. Теоретическая часть

### 2.1 Разработка алгоритма автоматического выделения СГ по термово-контекстному критерию из дерева зависимостей

В данном подпункте будет описана разработка алгоритмов для всех критериев, которые возможно реализовать.