**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ**

**им. А.И. ГЕРЦЕНА»**



Институт информационных технологий и технологического образования

Реферат на тему: «Обработка естественного языка»

Дисциплина: «Логика и методология науки»

студент группы 1А \_\_\_\_\_\_\_\_\_\_\_\_\_\_ / Момот Д.М.

профессор каф. ИС \_\_\_\_\_\_\_\_\_\_\_\_\_\_/ д.т.н., проф. Фомин В.В.

Санкт-Петербург

2020 г.

**Постановка задачи**. Компьютер должен понять текст, т. е. стать способным ответить на вопросы по нему (построить следствия?). Зачастую для понимания предложения на естественном языке необходимо не только понимание синтаксиса и лексики этого предложения, но и наличие цельной картины мира, позволяющей из значения отдельных слов составить смысл всего предложения со всеми его деталями (цель действия, обстоятельства, к чему относится прилагательные и местоимения).

**Нетривиальность задачи** состоит в наличии неоднозначности в естественном языке. Примеры:

* Вася опоздал на день рождения, потому что заходил в цветочный магазин. Зачем заходил?
* Бананы отдали обезьянам, потому что они были перезрелые/голодные. Кто они?

**Теоретической основой** данного направления является теория формальных грамматик (формальных языков). Смежными дисциплинами являются теория моделей., формальная семантика, компьютерная лингвистика. В свою очередь, эти дисциплины основываются на понятиях логики высказываний, логик первого и высших порядков, а также теории формальных систем.

**Основные определения**:

**Язык** – все множество правильно построенных текстов. Более строго – множество конечных слов (строк, цепочек) над конечным алфавитом, т. е. множество конечных последовательностей элементов алфавита.

Над языками определены операции, позволяющими определять новые языки на основе существующих.

* Конкатенация: слово нового языка есть конкатенация слова первого языка со словом из второго
* Пересечение: слово нового языка обязано присутствовать в обоих старых
* И другие

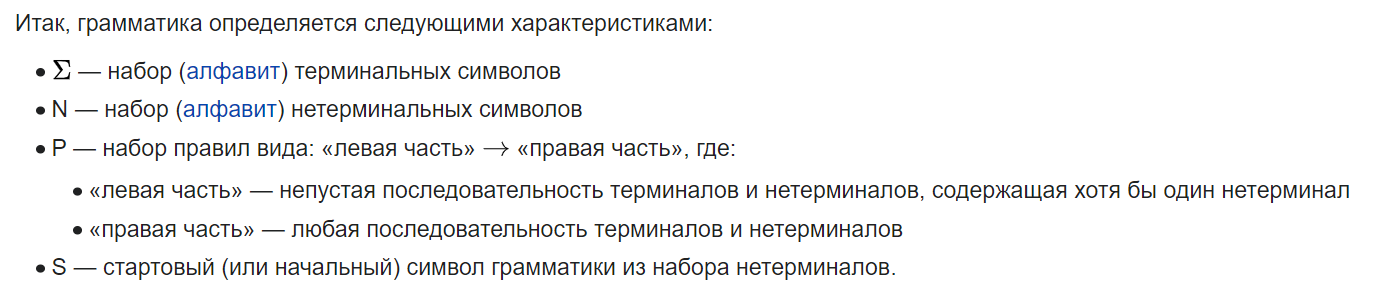
Язык может быть описан как:

1. Перечисление всех элементов (допустимых цепочек)
2. Слова, порожденные формальной грамматики (см ниже)
3. Слова, порожденные регулярным выражением – если язык является регулярным языком
4. Слова, распознаваемые конечным автоматом – если язык является автоматным языком
5. Слова, порожденные формой Бэкуса-Наура (?)

Согласно теореме Клини, множество регулярных языков совпадает с множеством автоматных, то есть определение языка регуляркой эквивалентно определению с помощью некоторого конечного автомата и наоборот.

Для описания языков удобно использовать **формальную грамматику** (или просто грамматику). По сути, она представляет собой набор правил вывода цепочек.

Грамматики бывают порождающие (задают правила порождения слов) и распознающие (аналитические) (по слову определяют, входит ли оно в язык).



Обычно стоит задача не построить все тексты, а распознать, правильно ли построено предложение, а потом желательно еще и разобрать его, построить дерево. Поэтому распознающие грамматики имеют больше приложений.

**Иерархия по Хомскому**. Каждый класс грамматик определяет, в свою очередь, соответствующий класс языков. И чем дальше, чем более богатым и сложным является язык.

* Регулярная (автоматная) – четкие правила, слева – всегда терминальный символ – однозначный разбор, можно использовать конечный автомат (или регулярку, помним про теорему Клини).
* Контекстно-свободная. Хотя нет однозначности, нет и зависимости от контекста (нет необходимости идти вперед) – регулярная грамматика с возможностью перебора. Можно строить дерево разбора. Широко применяются для синтаксиса языков программирования.
* Контекстно-зависимая – в правила входят фрагменты текста. Неплохо описывают предложения естественного языка, в компиляторах не используются из-за сложности. Для таких доказано, что по некоторому алгоритму можно за конечное число шагов определить, относится ли цепочка к языку.
* Рекурсивно-перечислимые (неограниченные). В силу сложности (неспецифицированности) практического применения не имеют.

**Приложения** данной задачи многочисленны и разнообразны. Ниже описываются приложения, относящиеся к анализу текста.

* Извлечение информации (information extracton) – извлечение структурированной информации из неструктурированных или слабоструктурированных цифровых документов. Примеры – автоматический сбор информации из новостей, контента соцсетей, бизнес-событий.
  + Распознавание имен собственных, в том числе названий
  + Поиск связей, относящихся к одному объекту
  + Выделение ключевых слов и словосочетаний для текста
  + Автореферирование (генеративное и декларативное) – выделение из текста смысловой, эмотивной, оценочной информации.
* Информационный поиск – поиск неструктурированной информации. Пример – поиск контента соцсетей, относящегося к заданной тематике. По методологии делится на адресный, семантический, документальный, фактографический.
  + Полнотекстовый – по всему тексту документа, так ищут интернет-поисковики. Обычно текст заранее индексируется, самый популярный – инвертированный индекс.
  + По метаданным – по атрибутам: название, дата создания, размер и т. д.
* Анализ высказываний
* Анализ тональности текста – автоматическое выделение эмоционально-окрашенной лексики и эмоциональной оценки (мнений). Тональность – эмоциональное отношение автора высказывания к объекту (объекту, событию, процессу или их свойствам), выраженное в тексте. На уровне лексемы – лексическая тональность, всего текста – функция лексических тональностей единиц текста (предложений). Классификация бывает по бинарной или многополосной шкале, а текже по субъективности/объективности. Могут применяться методы машинного обучения (с учителем или без), глубокого структурного разбора (теоретико-графовые) или использование тезаурусов (правила и словари).
* Вопросно-ответные системы.
  + Узкоспециализированные, например, медицина или обслуживание автомобилей. Естественно-языковый интерфейс БД.
  + Общие – работают по всем областям знаний, что позволяет искать ответы в смежных областях.

## Литература

Мотвани, Ульман, Хопкрофт: Введение в теорию автоматов, языков и вычислений.

Серебряков В. А., Галочкин М. П., Гончар Д. Р., Фуругян М. Г. Теория и реализация языков программирования

Кук Д., Бейз Г. Глава 8. Языки и грамматики

[Робин Хантер](https://ru.wikipedia.org/w/index.php?title=%D0%A0%D0%BE%D0%B1%D0%B8%D0%BD_%D0%A5%D0%B0%D0%BD%D1%82%D0%B5%D1%80&action=edit&redlink=1). Основные концепции компиляторов

Гладкий А. В. Формальные грамматики и языки

Кревский И. Г., Селивёрстов М. Н., Григорьева К. В. Формальные языки, грамматики и основы построения трансляторов: Учебное пособие

Мартыненко Б. К. Языки и трансляции: Учебное пособие

Пентус А. Е., Пентус М. Р. Математическая теория формальных языков

Фомичёв В. С. [Формальные языки, грамматики и автоматы: Курс лекций](http://old.eltech.ru/misc/LGA_2007_FINAL/Index.html) (https://old.etu.ru/misc/LGA\_2007\_FINAL/Index.html)

[Формализованный язык](https://iphlib.ru/library/collection/newphilenc/document/HASHbb6bbee180347b8a2096bb) / Б. В. Бирюков (https://iphlib.ru/library/collection/newphilenc/document/HASHbb6bbee180347b8a2096bb)

<https://edu.vsu.ru/pluginfile.php/80770/mod_resource/content/1/%D0%9E%D1%81%D0%BD%D0%BE%D0%B2%D1%8B%20%D1%81%D0%B8%D0%BD%D1%82%D0%B0%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B3%D0%BE%20%D1%80%D0%B0%D0%B7%D0%B1%D0%BE%D1%80%D0%B0.pdf>