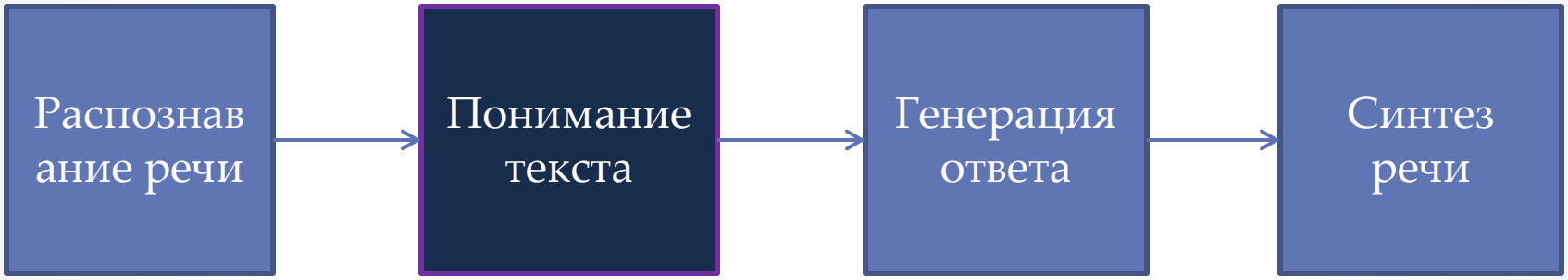
Разработка и программная реализация модуля для генерации достоверных гипотез пользовательского ввода в системах понимания естественной речи

Грицина Е.О., группа 43504/3

Обработка естественного языка

- **NLP** - Обработка естественного языка (Natural Language Processing)



- **NLU** – понимание естественной речи (Natural language understanding)

Извлечение информации из предложения

- Грамматические шаблоны и словари

USER SAYS

describe witness in case 1112

INTENT

DescribeWitness

ACTION

/private/search/cad/witness/describe

PARAMETER	VALUE
INCIDENT_ID	1112

Проблема распространенности предложений

- Не получится определить подходящий грамматический шаблон для предложений с распространяющими словами:

  - *Describe **the second** witness in case*
  - *Describe **female** witness in case*
  - *Describe witness **number two** in case*
  - *Describe **second or first** witness in case*
  - *Describe witness **in the last** case*

- Приходится добавлять новые грамматические шаблоны

Цели и задачи

- **Целью** работы является разработка алгоритма, который бы производил формирование гипотез пользовательского ввода на основе сказанного предложения и оценку достоверности этих гипотез.
- **Задачи:**
- Проанализировать существующие способы решения проблемы.
- Определить требования к алгоритму.
- Разработать алгоритм формирования гипотез.
- Разработать метод подсчета достоверности гипотез.
- Провести тестирование алгоритма и оценить результаты его работы.

Анализ существующих походов

- **Свободная позиция слов в предложении**

- Грамматический шаблон:
  * @describe * @witness *

- Будет верно применен к фразам:
✓ **Please** describe **for me second** witness **in case**
✓ Describe **female** witness **in last case**
✓ Describe **second or first** witness

- Но приведет к ложному срабатыванию для фраз:
✘ **Describe** vehicle of the **witness** in this case
✘ **Describe** incident with **witness**
✘ **Describe** route to the **witness**.

Анализ существующих походов

- **Алгоритмы суммаризации текстов**
- *Суммаризация текста представляет собой автоматическое выделение ключевой информации из текста и создание краткого изложения для него.*

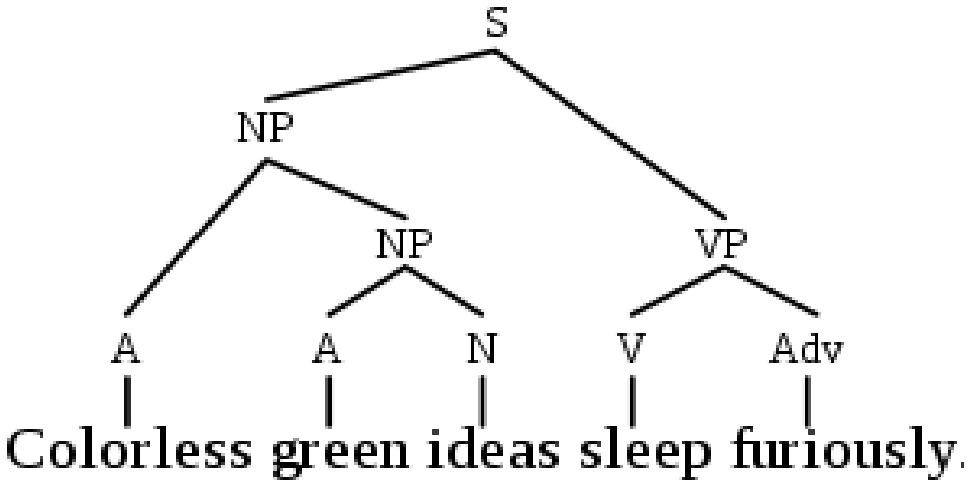
- Невозможно применять к предложениям, потому эти алгоритмы основаны на анализе связей между **несколькими предложениями текста**, выделению среди них ключевых и наиболее повторяющихся слов.

Требования к разрабатываемому алгоритму

- Формировать гипотезы на основе **исходной** фразы пользователя.
- Необходимо **исключить** из предложения распространяющие слова и конструкции.
- Изменение исходной фразы должно формировать новую гипотезу.
- **Достоверность** гипотезы должна изменяться в зависимости от количества исключенных слов, их частей речи и роли в предложении.
- В гипотезах необходимо сохранить **семантическое значение** исходной фразы
- Гипотезы должны обладать **синтаксической корректностью**.

Разработка алгоритма

- **Синтаксический анализ предложений**



- Принцип разделения на составляющие, каждая из которых тоже разбивается на составляющие.
- Будем формировать гипотезы на основе **анализа** таких синтаксических деревьев.

The Stanford Parser

- Синтаксический анализатор, входящий в библиотеку **Stanford CoreNLP**. Позволяет формировать синтаксические деревья и работать с ними.

Describe witness in case 12

```
(ROOT
  (S
    (VP (VB Describe)
      (NP (NN witness))
      (PP (IN in)
        (NP (NN case) (CD 12))))))
```

Правила семантического сокращения

- На основе анализа синтаксических деревьев было разработано 12 правил для формирования гипотез пользовательского ввода:

✓ **Пунктуация**
✓ Удаление и замена слова **Number**
✓ Притяжательное окончание **«'s»**
✓ **Прилагательные** перед существительным
✓ **Окончание «s»** в числительных
✓ Удаление **числительных**
✓ Удаление **наречий**
✓ Удаление **имен собственных**
✓ Удаление **равнозначных конструкций**
✓ 3 правила для работы с **предложными конструкциями**

Правила семантического сокращения

- **Удаление однородных конструкций**

What's the best way to get to witness one from here

```
(SBARQ
   (WHNP (WP what))
   (SQ (VBZ 's)
    (NP
     (NP (DT the) (JJS best) (NN way))
     (S
      (VP (TO to)
       (VP (VB get)
        (PP (TO to)
         (NP (NN witness) (CD one)))
        (PP (IN from)
         (NP (RB here)))))))))
```

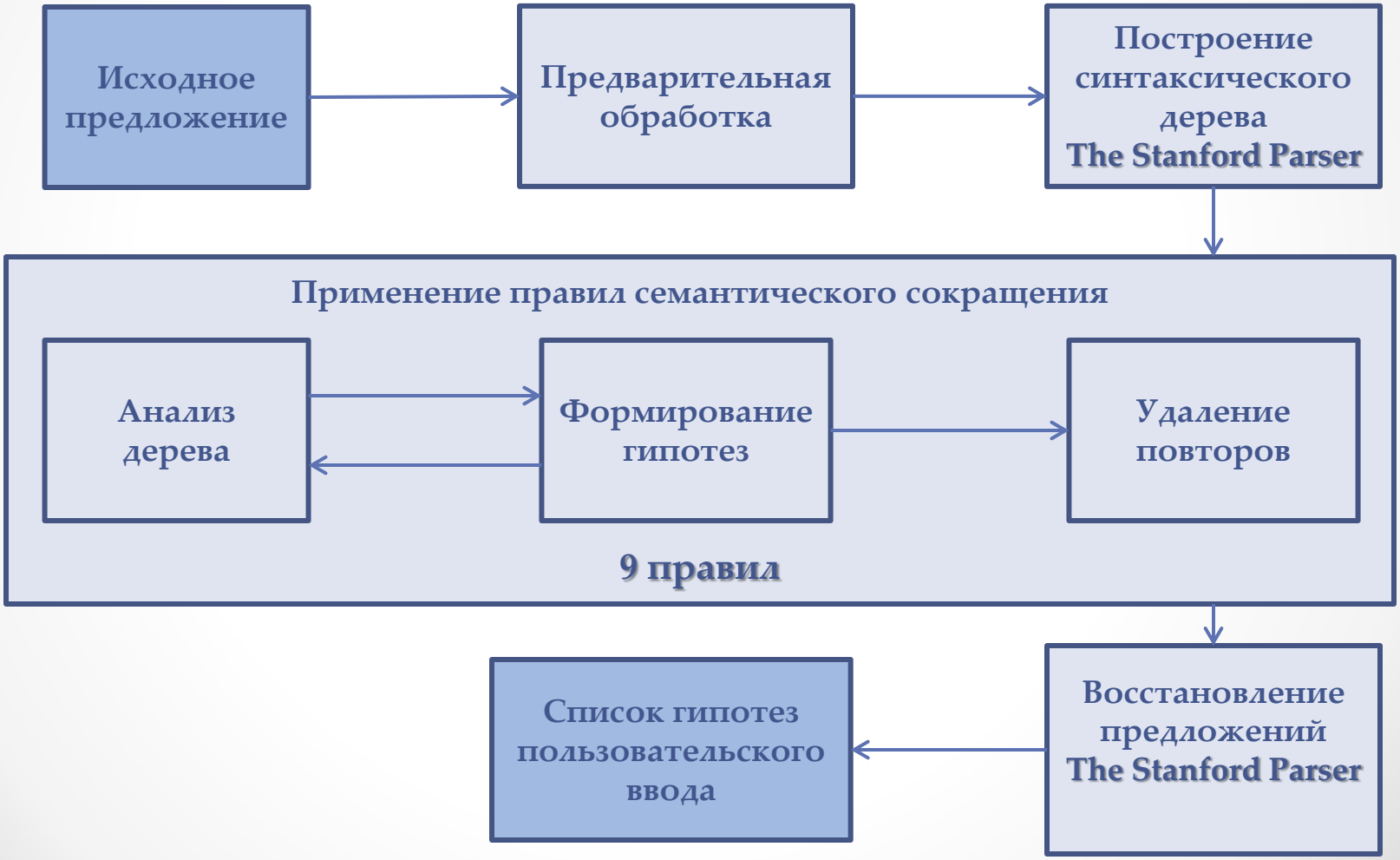
Правила семантического сокращения

- **Удаление предложных конструкций**

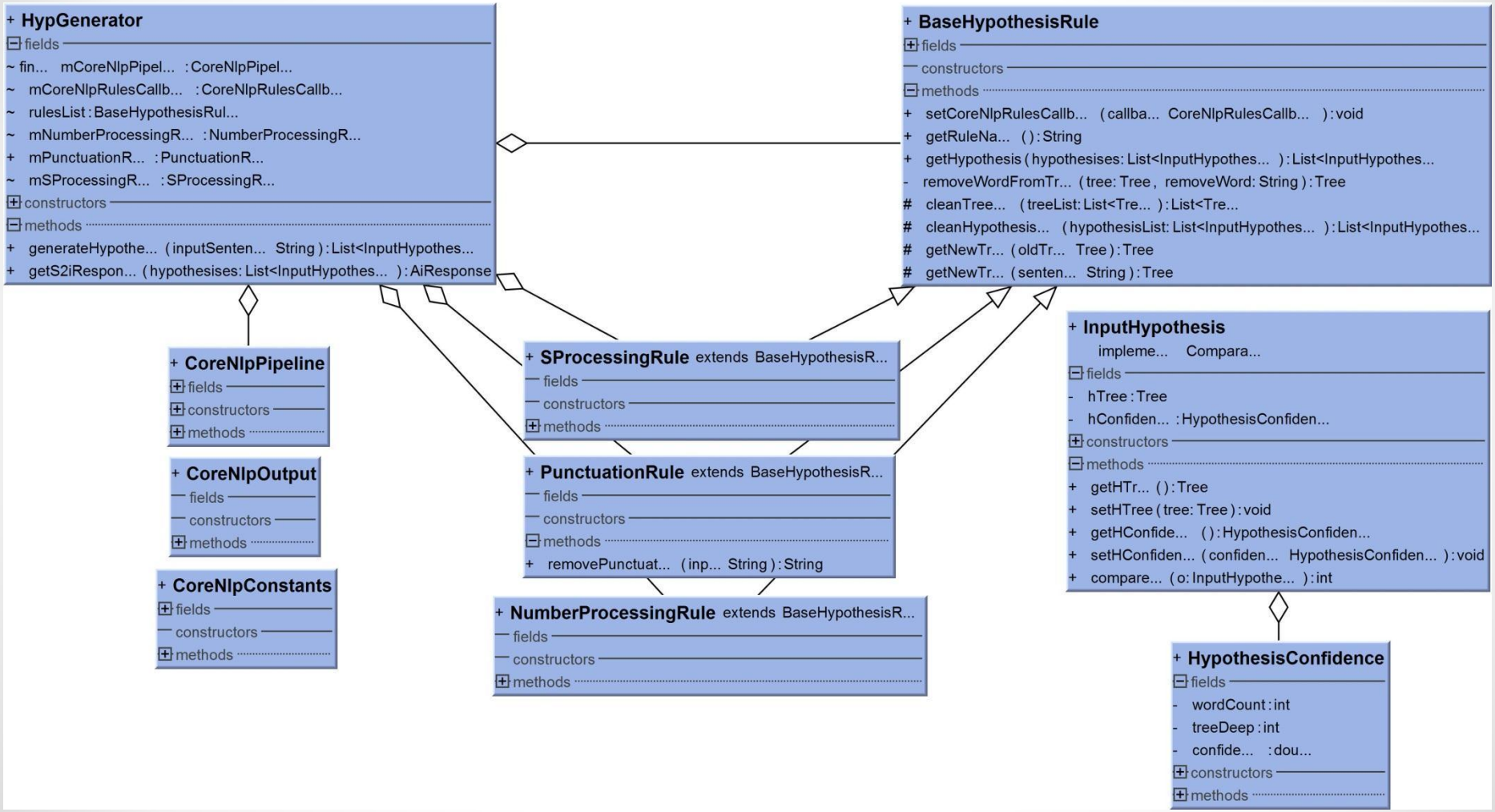
Find my video record with car crashing in the gallery

```
(VP (VB find)
    (NP (PRP$ my) (NN video) (NN record))
    (PP (IN with)
      (NP
        (NP (NN car))
        (VP (VBG crashing)
          (PP (IN in)
            (NP (DT the) (NN gallery)))))))))
```

Порядок формирования гипотез



Архитектура проекта



Оценка достоверности

- Происходит при формировании каждой гипотезы.
- Зависит от количества удаляемых слов и их частей речи.
- Каждая часть речи имеет свой «вес». Веса описаны в классе констант.
- Исходная фраза обладает максимальной достоверностью, гипотезы лишь уменьшают ее.
- Рассчитывается по формуле:

$$newConf = conf\ - conf * \left(\frac{rWc}{tWc}\right) * POSc * Rc;$$

- **newConf** и **conf** – новая и исходная достоверность, **rWc** – количество удаленных слов, **tWc** – количество слов до удаления, **POSc** и **Rc** – коэффициенты для частей речи и применяемого правила.

Результат работы алгоритма

Пример 1.

I need to get to witness 3 in case 8776

Ожидаемые гипотезы:

- I need to get to witness in case 8776
- I need to get to witness 3
- I need to get to witness in case
- I need to get to witness

Результат работы алгоритма:

1. c:1.0 : I need to get to witness 3 in case 8776
2. c:0.944 : I need to get to witness in case 8776
3. c:0.93 : I need to get to witness 3 in case
4. c:0.870 : I need to get to witness in case
5. c:0.803 : I need to get to witness 3
6. c:0.739 : I need to get to witness

Сформированные гипотезы: 4 из 4

Результат работы алгоритма

Пример 2.

let make our way to the residence of witness number 2

Ожидаемые гипотезы:

- let make way to the residence of witness 2
- let make way to the residence of witness
- let make way to witness
- let make way to the residence

Результат работы алгоритма:

1. c:0.981 : let make our way to the residence of witness 2
2. c:0.942 : let make way to the residence of witness 2
3. c:0.913 : let make our way to the residence of witness
4. c:0.869 : let make way to the residence of witness
5. c:0.788 : let make our way to the residence
6. c:0.738 : let make way to the residence

Сформированные гипотезы: 3 из 4

Тестирование алгоритма

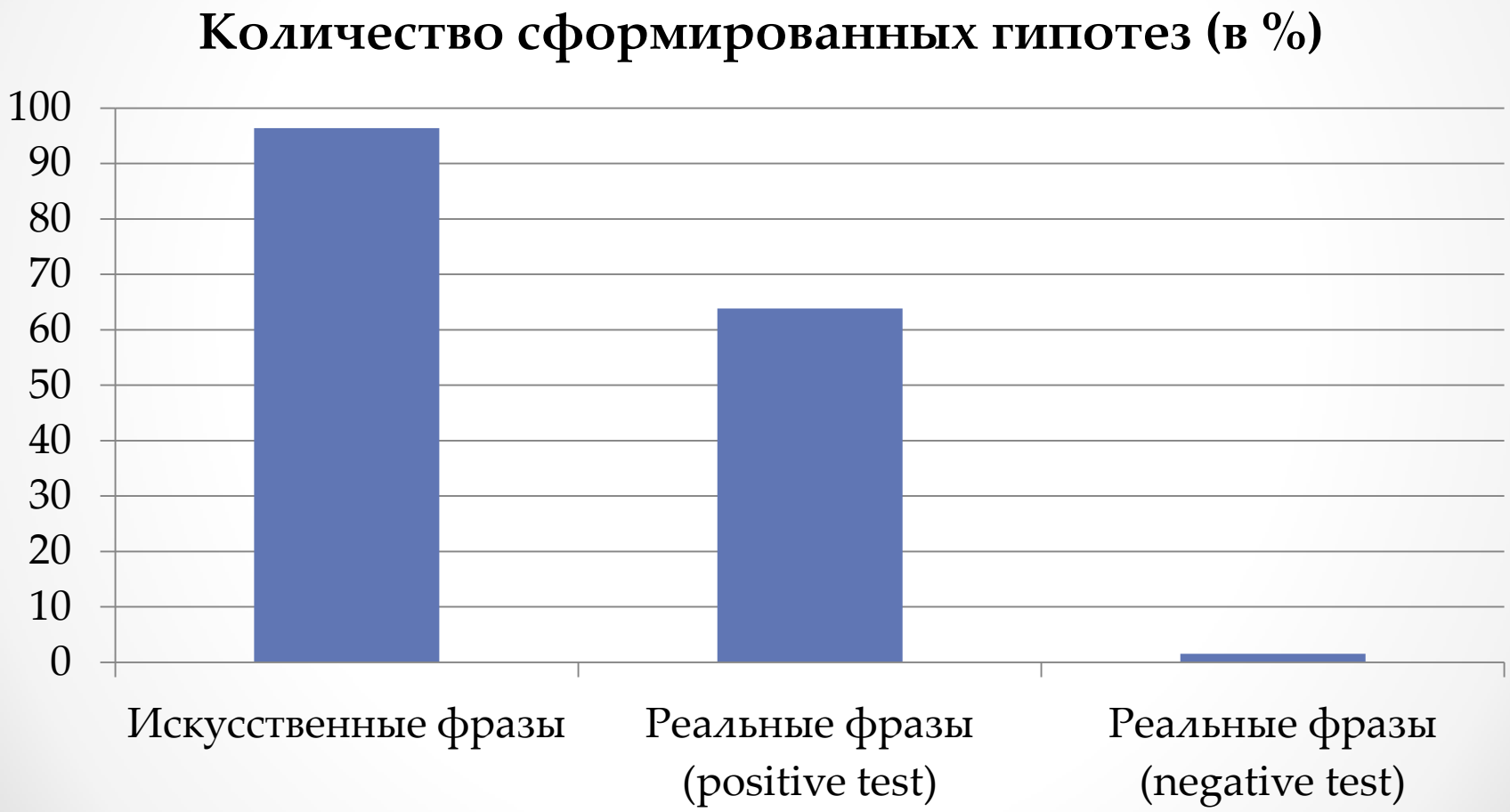
- **Три критерия оценки работы алгоритма:**

1. Отношение количества сформированных гипотез к ожидаемым для искусственных фраз.
2. Отношение количества сформированных гипотез к ожидаемым для реальных запросов пользователей.
3. Отношение количества сформированных гипотез к нежелательным гипотезам реальных запросов пользователей.

Тестирование алгоритма

Тестовый набор	Количество ожидаемых гипотез	Удалось сгенерировать	Кол-во фраз, для которых сработал алгоритм
Искусственные фразы	195	188	49 из 50
Реальные фразы (positive test)	83	53	23 из 26
Реальные фразы (negative test)	198	3	3 из 182

Оценка результатов работы алгоритма



Заключение

- Интеграция данного алгоритма в существующую систему понимания естественной речи позволила бы существенно повысить точность работы такой системы без необходимости изменения существующих грамматических шаблонов или создания новых.
- Таким образом, все поставленные задачи в ходе работы были выполнены, а цель работы можно считать достигнутой.

Спасибо за внимание!