

# Математические модели в лингвистике

## Вводная лекция, часть 2.

Мати Пентус, Александр Пиперски, Алексей Сорокин

МГУ им. М. В. Ломоносова, межфакультетский курс,  
осенний семестр 2017–2018 учебного года,  
12 сентября 2017 г.

# Разделы лингвистики

- Разделы лингвистики:
  - Морфология,
  - Синтаксис,
  - Фонетика,
  - Семантика,
  - Корпусная лингвистика,
  - ...,
- Как в них применяется математика?
- И зачем нужны они сами в нелингвистических задачах?

# Информационный поиск

- Одно из применений математической лингвистики — интернет-поиск.
- Казалось бы, лингвистика в поиске не нужна (просто ищем совпадение...),
- Однако всё не так просто:
  - Нужны не просто совпадения, а совпадения слов,
  - Слова могут изменяться по грамматическим категориям,
  - Пользователь может делать опечатки,
  - Иногда пользователь не совсем знает, что ему нужно (стоит выдавать синонимы...),
  - Слова могут быть многозначны.

# Пример поискового запроса

The screenshot shows a Google search results page for the query "недостаток метод". The search bar contains the text "недостаток метод". The results are sorted by "Все" (All). The first result is "Недостаток - метод - Большая Энциклопедия Нефти и Газа, статья ...". The second result is "Метод Дельфи - Центр креативных технологий". The third result is "Метод Бринелля — Википедия". The fourth result is "Дихотомия — Википедия". The fifth result is "Достоинства и недостатки методов оценки инвестиционных ...". The sixth result is "Метод симуляций. Основной недостаток метода симуляций — ..." (partially visible).

Результатов: примерно 17 400 000 (0,74 сек.)

**Недостаток - метод** - Большая Энциклопедия Нефти и Газа, статья ...  
[www.ngpedia.ru/id162684p2.html](http://www.ngpedia.ru/id162684p2.html)  
 Недостаток метода заключается в необходимости через каждые шесть месяцев ( при среднем содержании в исходной воде до 10 мг / л SiO2) менять ...

**Метод Дельфи** - Центр креативных технологий  
<https://inventech.ru/pub/methods/metod-0013/>  
 Авторы метода: О. Холмер, Т. Гордон и др. (США), 50-е годы ... Метод Дельфи - один из инструментов выбора и оценки решения. ... Недостатки метода.

**Метод Бринелля** — Википедия  
[https://ru.wikipedia.org/wiki/Метод\\_Бринелля](https://ru.wikipedia.org/wiki/Метод_Бринелля)  
 Перейти к разделу **Преимущества** и **недостатки** - Недостатки. Метод рекомендуется применять для материалов с твердостью до 450 НВ.

**Дихотомия** — Википедия  
<https://ru.wikipedia.org/wiki/Дихотомия>  
 Дихотомия (греч. διχοτύπος букл. «надвое» + тожд. «деление») — разделение, ... Дихотомическое деление имеет недостаток: при делении объема ... Рассмотрим метод дихотомии условной одномерной оптимизации (для ...


**Достоинства и недостатки методов оценки инвестиционных ...**  
[gaap.ru/\\_idostoinstva\\_i\\_nedostatki\\_metodov\\_otsenki\\_investitsionnykh\\_proektov/](http://gaap.ru/_idostoinstva_i_nedostatki_metodov_otsenki_investitsionnykh_proektov/)  
 16 сент. 2005 г. - Достоинства и недостатки методов оценки инвестиционных проектов. Если вы собственник или топ-менеджер организации, то вам, ...


**Метод симуляций. Основной недостаток метода симуляций — ...**


## Выводы из примера


- Может быть найдено не только слово, но его форма (**метод**  $\mapsto$  **метода**),
- Может быть найдено близкое по смыслу слово (**метод**  $\mapsto$  **методика**),

Надежный | <https://www.google.ru/search?num=100&newwindow=1&safe=off&q=недостаток+м>

Яндекс  Почта

[PDF](#) Терехов А.Н. Недостаток интрузивного метода оценки качества ...   
[conf.mirea.ru/CD2016/pdf/p5/22.pdf](http://conf.mirea.ru/CD2016/pdf/p5/22.pdf) ▼  
 25 нояб. 2016 г. - методы оценки качества передачи речи обладают высокой точностью и ...  
 Недостаток интрузивных методов - увеличение тра-

Методика решения задач на «избыток–недостаток» в курсе ...   
<https://him.1september.ru/2003/44/25.htm> ▼  
 Методика решения задач на «избыток–недостаток» в курсе основной общеобразовательной  
 школы. Умение решать химические задачи – важная ...

Метод электрокоагуляции, преимущества и недостатки — БУЗВО ...   
[vokkvd.ru/patients/метод-электрокоагуляции-преимущест/](http://vokkvd.ru/patients/метод-электрокоагуляции-преимущест/) ▼  
 Метод электрокоагуляции, преимущества и недостатки. Возможности метода: удаление  
 остроконечный кондилом, папиллом, бородавок, контактного ...

## Выводы из примера

- Может быть найдено не только слово, но его форма (**метод**  $\mapsto$  **метода**),
- Может быть найдено близкое по смыслу слово (**метод**  $\mapsto$  **методика**),
- Может быть исправлена опечатка (**недостарок**  $\mapsto$  **недостаток**),

Google

Все Картинки Видео Новости Карты Ещё Настройки Инструменты

Результатов: примерно 848 000 (0,90 сек.)

Показаны результаты по запросу **недостаток** метод  
Искать вместо этого **недостарок** метод

**Метод Дельфи - Центр креативных технологий**  
<https://inventech.ru/pub/methods/metod-0013/>  
 Авторы метода: О. Холмер, Т. Гордон и др. (США). 50-е годы ... Метод Дельфи - один из инструментов выбора и оценки решения. ... Недостатки метода

**Каковы основные недостатки абсорбц и адсорбц. методов ...**  
<https://all-ecology.ru/index.php?request=full&id=119>  
 Существенным недостатком сорбционных методов очистки (абсорбционных и адсорбционных) выбросных газов является необходимость ...

**Достоинства и недостатки индуктивных методов познания**  
<https://www.filosofia.ru/...metod.../665-dostoinstva-i-nedostatki-induktivnyh-metodo...>  
 Достоинства и недостатки индуктивных методов познания. У индукции имеется серьезное преимущество перед дедукцией - она основывается на ...

**Метод Бринелля — Википедия**  
[https://ru.wikipedia.org/wiki/Метод\\_Бринелля](https://ru.wikipedia.org/wiki/Метод_Бринелля)  
 Перейти к разделу **Преимущества** и **недостатки** - Недостатки. Метод рекомендуется применять для материалов с твердостью до 450 НВ.

# Вычислительная морфология

- Одна из задач вычислительной морфологии: восстановление форм слова по базовой форме,

word+PI  $\mapsto$  words,

state+PI  $\mapsto$  ?,

word+PI  $\mapsto$  words,

state+PI  $\mapsto$  states,

- Однако не всё так просто:

torch+PI  $\mapsto$  torches,

daisy+PI  $\mapsto$  daisies,

array+PI  $\mapsto$  arrays,

mouse+PI  $\mapsto$  mice,

goose+PI  $\mapsto$  geese,

cactus+PI  $\mapsto$  cacti, cactuses,

# Код для английского языка

```

### english.foma ###

read lexc irregular.lexc
define IrregularNounPlural;

define Vowel [ a | i | e | o | u | y ];
define Consonant [ b | c | d | f | g | h | j | k | l | m | n | p | q | r | s | t |
  v | w | x | z ];
define Letter [ Vowel | Consonant ];
define Word [ Letter ]+;
define NounMark "+N";
define NounNumber "+Sg" | "+Pl";
define Noun Word NounMark NounNumber;

define NounAffixation "+N" "+Sg" -> "^" || _ .#., "+N" "+Pl" -> "^" s || _ .#.;
define yReplacement y -> i e || Consonant _ "^" s .#.;
define sibException [ Letter ]+ a r c h "^" s ;
define Sibilant [ x | s | z | c h | s h ];
define eInsertion [..] -> e || Sibilant _ "^" s .#.;
define checkSibilant [ sibException .P. eInsertion ];
define Cleanup "^" -> [] || _ ;
define RegularNoun [ NounAffixation .o. yReplacement .o. checkSibilant .o. Cleanup
  ] ;
define Grammar Noun .o. [ IrregularNounPlural .P. RegularNoun ] ;

push Grammar

```



# Русская глагольная морфология

Для русского всё куда хуже:

<i>играть</i>	↦	<i>играет</i>
<i>писать</i>	↦	<i>пишет</i>
<i>спать</i>	↦	<i>спит</i>
<i>греть</i>	↦	<i>греет</i>
<i>хрипеть</i>	↦	<i>хрипит</i>
<i>петь</i>	↦	<i>поёт</i>
<i>целовать</i>	↦	<i>целует</i>

## Глагольные формы в языке йоулумни

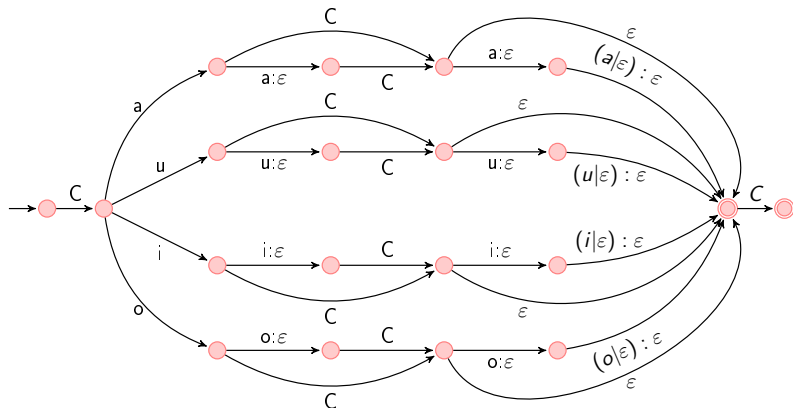
основа	герундий	дуратив
сaw “кричать”	сaw-inay	сawaa-ʔaa-n
сuиm “разрушать”	сuиm-inay	сuиии-ʔaa-n
hoyoo “называть”	hoy-inay	hoyoo-ʔaa-n
diiyl “охранять”	diyl-inay	diyiiil-ʔaa-n
ʔilk “петь”	ʔilk-inay	ʔiliik-ʔaa-n
hiwiit “гулять”	hiwt-inay	hiwiit-ʔaa-n

## Глагольные формы в языке йоулумни

Если основа имела вид  $\alpha_1 V(V)\alpha_2(V)(V)\alpha_3$ , где  $\alpha_1, \alpha_2, \alpha_3 \in \{C, \varepsilon\}$ , то основа герундия имеет вид  $\alpha_1 V\alpha_2\alpha_3$ , а основа дуратива —  $\alpha_1 V\alpha_2 VV\alpha_3$ .

## Преобразователи для глагольных форм в языке йоулумни

Основа герундия:



# Постановка задачи

- Текстовая классификация — задача автоматического распределения текстов по классам.
- Классы зависят от задачи:
  - Анализ тональности: положительные, отрицательные и нейтральные отзывы.
  - Жанровая классификация: тексты различных жанров (новостные, энциклопедические, художественные...).
  - Тематическая классификация: распределение текстов по темам.
  - Языковая классификация: автоматический анализ языка.
  - Можно классифицировать по полу, возрасту, социальному положению автора...
- Применения:
  - Информационный поиск,
  - Автоматическая рубрикация,
  - Рекомендационные системы.
  - Автоматическое определение авторства.

## Модель “мешка слов”

- Для каждого слова учитывается только число его вхождений в текст.
- Порядок слов не играет роли.
- Получаем матрицу (таблицу) “объекты-признаки”.
- Документ — строка матрицы (вектор значений признаков).

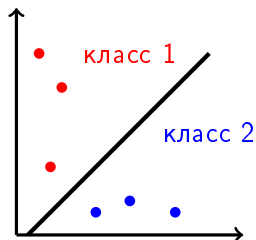
документ	office	web	...	learn	метка класса
1	2	2	...	0	student
5	0	2	...	3	project
48	2	4	...	1	course

Пример матрицы “объекты-признаки” (WebKB corpus, McCallum, 1998)

- Метод простой, но он работает во многих задачах.

## Особенности текстовой классификации

- Разделяющая поверхность — гиперплоскость



- Классификатор можно задать формулой:

$$h(x) = \operatorname{sgn} \left( \sum_{i=1}^n w_i x_i - w_0 \right), \quad x_i \text{ — компоненты вектора } x$$

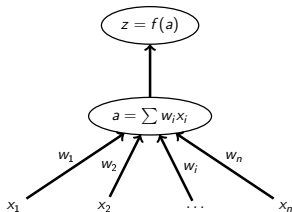
- Веса  $w_0, w_1, \dots, w_n$  подбираются автоматически по обучающей выборке (текстам, для которых известна метка класса).

## Особенности текстовой классификации

- Признаки могут учитывать контекст (биграммы, триграммы и т.д.),
- В качестве признаков можно использовать длину предложения, статистику по грамматическим категориям, синтаксические показатели...
- Подбор признаков — задача лингвиста.
- Оптимальная классификация признаков — задача математика.
- Трудности: большой объём данных, несбалансированная выборка, большая размерность данных, большое количество классов.

# Нейронные сети

- Один из основных подходов: нейронные сети.
- Структура нейрона:

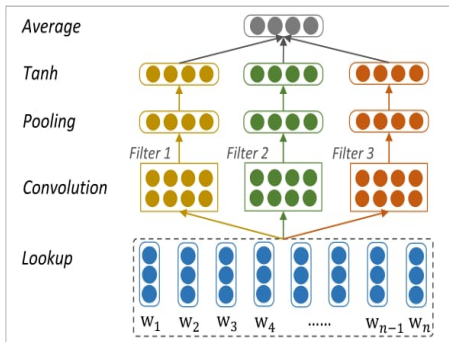


- Нейрон получает несколько сигналов  $x_1, \dots, x_n$  и вычисляет их взвешенную сумму.
- Эта сумма поступает в функцию активации  $f$ .
- Варианты для функции  $f$ :
  - $f(x) = x$  (тождественная).
  - $f(x) = \text{sgn}(x)$  (преодолен ли порог активации).
  - $f(x) = \frac{1}{1 + e^{-kx}}$  (преобразует  $(-\infty; \infty)$  в  $[0; 1]$ ).
  - $f(x) = \max(x - b, 0)$  (ReLU, смесь тождественной и пороговой).



# Нейронные сети

Реальные нейронные сети: десятки тысяч нейронов, организованных в слои.



Структура сети и признаки зависят от лингвистической задачи (но не так сильно, как при классификации).