

Лингвистика и формальные языки

Марина Ермолаева

Математические методы в лингвистических исследованиях

8 февраля 2022

- Математические методы в лингвистических исследованиях

- Математические методы в лингвистических исследованиях
(Maybe Remote But Hopefully In-Person Edition)

- Математические методы в лингвистических исследованиях
(Maybe Remote But Hopefully In-Person Edition)

- Марина Борисовна Ермолаева

- Электронная почта
 - mail@mermolaeva.com
 - Объявления, домашние задания, etc.
 - Отвечаю на письма в течение 24 часов
- Dropbox
 - Общая папка: [ссылка](#)
 - Здесь будут материалы к курсу
- Анонимные вопросы и комментарии
 - Google Forms: [ссылка](#)

- Чат
 - для коротких комментариев

- Невербальная обратная связь



- Аудио + видео
 - для вопросов и комментариев; сначала поднимите руку

- Каждые ≈ 2 недели
- Присылать на электронную почту **в формате PDF**
(использование \LaTeX поощряется)
- Если не указано иное, задания **можно** обсуждать друг с другом!

При этом необходимо:

- Указать, с кем обсуждали
- Описать ход решения лично своими словами

- В этом семестре **зачет**, в следующем **экзамен**
- Компоненты зачета:
 - Домашние задания
 - Активность в классе

Математика в лингвистике? (“Ф” или “П” в ФиПЛ?)

Математика в лингвистике? (“Ф” или “П” в ФиПЛ?)

- Вычислительная лингвистика
(computational linguistics)
– гипероним: всё и сразу


Математика в лингвистике? (“Ф” или “П” в ФиПЛ?)

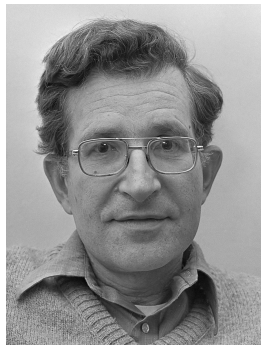
- Вычислительная лингвистика
(computational linguistics)
– гипероним: всё и сразу
- Автоматическая обработка естественного языка
(natural language processing)
– решение практических задач, связанных с естественным языком

Математика в лингвистике? (“Ф” или “П” в ФиПЛ?)

- Вычислительная лингвистика
(computational linguistics)
– гипероним: всё и сразу
- Автоматическая обработка естественного языка
(natural language processing)
– решение практических задач, связанных с естественным языком
- Математическая лингвистика
(mathematical linguistics)
– описание естественного языка математическими методами
– **теория формальных языков**: языки как математические объекты, порождаемые системами правил

Математика в лингвистике? (“Ф” или “П” в ФиПЛ?)

- Вычислительная лингвистика
(computational linguistics)
– гипероним: всё и сразу
- Автоматическая обработка естественного языка
(natural language processing)
– решение практических задач, связанных с естественным языком
- Математическая лингвистика
(mathematical linguistics) 
– описание естественного языка математическими методами
– **теория формальных языков**: языки как математические объекты, порождаемые системами правил



“Поиски строгих формулировок в лингвистике вызываются гораздо более серьезными мотивами, чем просто желанием соблюсти логические тонкости или упорядочить традиционные методы лингвистического анализа. [...] Выводя неприемлемые следствия из точных, но неадекватных формулировок, мы часто можем с большой точностью установить причину этой неадекватности и, таким образом, получить более глубокое представление о лингвистических данных.”

Ноам Хомский, *Синтаксические структуры*
(1957)

- **Алфавит:** конечное множество символов; обычно обозначается Σ

- **Алфавит:** конечное множество символов; обычно обозначается Σ
- Примеры:

- **Алфавит:** конечное множество символов; обычно обозначается Σ
- Примеры:
 - $\{a, б, в, \dots, я\}$ – алфавит

- **Алфавит:** конечное множество символов; обычно обозначается Σ
- Примеры:
 - $\{a, б, в, \dots, я\}$ – алфавит
 - $\{0, 1, +, -, =\}$ – алфавит

- **Алфавит:** конечное множество символов; обычно обозначается Σ
- Примеры:
 - $\{a, б, в, \dots, я\}$ – алфавит
 - $\{0, 1, +, -, =\}$ – алфавит
 - $\{\text{t}\}$

- **Алфавит:** конечное множество символов; обычно обозначается Σ
- Примеры:
 - $\{a, б, в, \dots, я\}$ – алфавит
 - $\{0, 1, +, -, =\}$ – алфавит
 - $\{\text{б}\}$ – алфавит

- **Алфавит:** конечное множество символов; обычно обозначается Σ
- Примеры:
 - $\{a, б, в, \dots, я\}$ – алфавит
 - $\{0, 1, +, -, =\}$ – алфавит
 - $\{\text{б}\}$ – алфавит
 - $\{и, в, не, на, я\}$, пять самых частотных слов русского языка по версии Викисловаря

- **Алфавит:** конечное множество символов; обычно обозначается Σ
- Примеры:
 - $\{a, б, в, \dots, я\}$ – алфавит
 - $\{0, 1, +, -, =\}$ – алфавит
 - $\{\text{б}\}$ – алфавит
 - $\{и, в, не, на, я\}$, пять самых частотных слов русского языка по версии Викисловаря – алфавит

- **Алфавит:** конечное множество символов; обычно обозначается Σ
- Примеры:
 - $\{a, б, в, \dots, я\}$ – алфавит
 - $\{0, 1, +, -, =\}$ – алфавит
 - $\{\text{б}\}$ – алфавит
 - $\{и, в, не, на, я\}$, пять самых частотных слов русского языка по версии Викисловаря – алфавит
 - $\mathbb{N} = \{0, 1, 2, 3, \dots\}$, множество натуральных чисел

- **Алфавит:** конечное множество символов; обычно обозначается Σ
- Примеры:
 - $\{a, б, в, \dots, я\}$ – алфавит
 - $\{0, 1, +, -, =\}$ – алфавит
 - $\{\text{б}\}$ – алфавит
 - $\{и, в, не, на, я\}$, пять самых частотных слов русского языка по версии Викисловаря – алфавит
 - $\mathbb{N} = \{0, 1, 2, 3, \dots\}$, множество натуральных чисел – **не** алфавит

- **Алфавит:** конечное множество символов; обычно обозначается Σ
- Примеры:
 - $\{a, б, в, \dots, я\}$ – алфавит
 - $\{0, 1, +, -, =\}$ – алфавит
 - $\{\text{б}\}$ – алфавит
 - $\{и, в, не, на, я\}$, пять самых частотных слов русского языка по версии Викисловаря – алфавит
 - $\mathbb{N} = \{0, 1, 2, 3, \dots\}$, множество натуральных чисел – **не** алфавит
 - Множество всех грамматичных предложений русского языка

- **Алфавит:** конечное множество символов; обычно обозначается Σ
- Примеры:
 - $\{a, б, в, \dots, я\}$ – алфавит
 - $\{0, 1, +, -, =\}$ – алфавит
 - $\{\text{б}\}$ – алфавит
 - $\{и, в, не, на, я\}$, пять самых частотных слов русского языка по версии Викисловаря – алфавит
 - $\mathbb{N} = \{0, 1, 2, 3, \dots\}$, множество натуральных чисел – **не** алфавит
 - Множество всех грамматичных предложений русского языка – **не** алфавит

- **Строка (или слово)** в алфавите Σ :
конечная последовательность символов из Σ

- **Строка (или слово)** в алфавите Σ :
конечная последовательность символов из Σ
- Пустая строка обозначается ϵ

- **Строка (или слово)** в алфавите Σ :
конечная последовательность символов из Σ
- Пустая строка обозначается ϵ
- Σ^* : множество всех строк в алфавите Σ
Если $\Sigma = \{a, b\}$, то $\Sigma^* = \{\epsilon, a, b, aa, ab, ba, bb, aaa, \dots\}$

- **Строка (или слово)** в алфавите Σ :
конечная последовательность символов из Σ
- Пустая строка обозначается ϵ
- Σ^* : множество всех строк в алфавите Σ
Если $\Sigma = \{a, b\}$, то $\Sigma^* = \{\epsilon, a, b, aa, ab, ba, bb, aaa, \dots\}$
- Σ^+ : множество всех непустых строк в алфавите Σ

- **Строка (или слово)** в алфавите Σ :
конечная последовательность символов из Σ
- Пустая строка обозначается ϵ
- Σ^* : множество всех строк в алфавите Σ
Если $\Sigma = \{a, b\}$, то $\Sigma^* = \{\epsilon, a, b, aa, ab, ba, bb, aaa, \dots\}$
- Σ^+ : множество всех непустых строк в алфавите Σ
- Если x – строка и $n \in \mathbb{N}$, то x^n – слово x , повторенное n раз

- **Строка (или слово)** в алфавите Σ :
конечная последовательность символов из Σ
 - Пустая строка обозначается ϵ
 - Σ^* : множество всех строк в алфавите Σ
Если $\Sigma = \{a, b\}$, то $\Sigma^* = \{\epsilon, a, b, aa, ab, ba, bb, aaa, \dots\}$
 - Σ^+ : множество всех непустых строк в алфавите Σ
 - Если x – строка и $n \in \mathbb{N}$, то x^n – слово x , повторенное n раз
- Примеры:

- **Строка (или слово)** в алфавите Σ :
конечная последовательность символов из Σ
- Пустая строка обозначается ϵ
- Σ^* : множество всех строк в алфавите Σ
Если $\Sigma = \{a, b\}$, то $\Sigma^* = \{\epsilon, a, b, aa, ab, ba, bb, aaa, \dots\}$
- Σ^+ : множество всех непустых строк в алфавите Σ
- Если x – строка и $n \in \mathbb{N}$, то x^n – слово x , повторенное n раз

Примеры:

- $a^3 = aaa$

- **Строка** (или **слово**) в алфавите Σ :
конечная последовательность символов из Σ
- Пустая строка обозначается ϵ
- Σ^* : множество всех строк в алфавите Σ
Если $\Sigma = \{a, b\}$, то $\Sigma^* = \{\epsilon, a, b, aa, ab, ba, bb, aaa, \dots\}$
- Σ^+ : множество всех непустых строк в алфавите Σ
- Если x – строка и $n \in \mathbb{N}$, то x^n – слово x , повторенное n раз

Примеры:

- $a^3 = aaa$
- $ab^2 = abb$

- **Строка (или слово)** в алфавите Σ :
конечная последовательность символов из Σ
- Пустая строка обозначается ϵ
- Σ^* : множество всех строк в алфавите Σ
Если $\Sigma = \{a, b\}$, то $\Sigma^* = \{\epsilon, a, b, aa, ab, ba, bb, aaa, \dots\}$
- Σ^+ : множество всех непустых строк в алфавите Σ
- Если x – строка и $n \in \mathbb{N}$, то x^n – слово x , повторенное n раз

Примеры:

- $a^3 = aaa$
- $ab^2 = abb$
- $(ab)^2 =$

- **Строка** (или **слово**) в алфавите Σ :
конечная последовательность символов из Σ
- Пустая строка обозначается ϵ
- Σ^* : множество всех строк в алфавите Σ
Если $\Sigma = \{a, b\}$, то $\Sigma^* = \{\epsilon, a, b, aa, ab, ba, bb, aaa, \dots\}$
- Σ^+ : множество всех непустых строк в алфавите Σ
- Если x – строка и $n \in \mathbb{N}$, то x^n – слово x , повторенное n раз

Примеры:

- $a^3 = aaa$
- $ab^2 = abb$
- $(ab)^2 = abab$

- **Строка** (или **слово**) в алфавите Σ :
конечная последовательность символов из Σ
- Пустая строка обозначается ϵ
- Σ^* : множество всех строк в алфавите Σ
Если $\Sigma = \{a, b\}$, то $\Sigma^* = \{\epsilon, a, b, aa, ab, ba, bb, aaa, \dots\}$
- Σ^+ : множество всех непустых строк в алфавите Σ
- Если x – строка и $n \in \mathbb{N}$, то x^n – слово x , повторенное n раз

Примеры:

- $a^3 = aaa$
- $ab^2 = abb$
- $(ab)^2 = abab$
- $b^0 =$

- **Строка** (или **слово**) в алфавите Σ :
конечная последовательность символов из Σ
- Пустая строка обозначается ϵ
- Σ^* : множество всех строк в алфавите Σ
Если $\Sigma = \{a, b\}$, то $\Sigma^* = \{\epsilon, a, b, aa, ab, ba, bb, aaa, \dots\}$
- Σ^+ : множество всех непустых строк в алфавите Σ
- Если x – строка и $n \in \mathbb{N}$, то x^n – слово x , повторенное n раз

Примеры:

- $a^3 = aaa$
- $ab^2 = abb$
- $(ab)^2 = abab$
- $b^0 = \epsilon$

- **Язык** над алфавитом Σ :
(конечное или бесконечное) подмножество Σ^*

(Формальные) языки

- Язык над алфавитом Σ :
(конечное или бесконечное) подмножество Σ^*
- Примеры языков над алфавитом $\Sigma = \{a, b\}$:

(Формальные) языки

- **Язык** над алфавитом Σ :
(конечное или бесконечное) подмножество Σ^*
- Примеры языков над алфавитом $\Sigma = \{a, b\}$:
 - \emptyset
 - $\{\epsilon\}$
 - Σ
 - Σ^*
 - все последовательности из a и b короче 10 символов
 - все последовательности из a четной длины
 - ...

Пример: квадратные скобки

- Алфавит $\Sigma = \{ [,] \}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$

Пример: квадратные скобки

- Алфавит $\Sigma = \{ [,] \}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

Пример: квадратные скобки

- Алфавит $\Sigma = \{ [,] \}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

$[] []$

Пример: квадратные скобки

- Алфавит $\Sigma = \{ [,] \}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

$[] []$ (да)

Пример: квадратные скобки

- Алфавит $\Sigma = \{[,]\}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

[] [] (да)

[[]]

Пример: квадратные скобки

- Алфавит $\Sigma = \{ [,] \}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?
 - $[] []$ (да)
 - $[[]]$ (да)

Пример: квадратные скобки

- Алфавит $\Sigma = \{[,]\}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

[] [] (да)

[[]] (да)

][][]

Пример: квадратные скобки

- Алфавит $\Sigma = \{ [,] \}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

$[] []$ (да)

$[[]]$ (да)

$] [] [$ (нет)

Пример: квадратные скобки

- Алфавит $\Sigma = \{[,]\}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

[] [] (да)

[[]] (да)

][][] (нет)

[[][]]

Пример: квадратные скобки

- Алфавит $\Sigma = \{ [,] \}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

$[] []$ (да)

$[[]]$ (да)

$] [] [$ (нет)

$[[[]]]$ (да)

Пример: квадратные скобки

- Алфавит $\Sigma = \{[,]\}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

[] [] (да)

[[]] (да)

][][] (нет)

[[][]] (да)

[[][]][[]]

Пример: квадратные скобки

- Алфавит $\Sigma = \{[,]\}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

$[] []$ (да)

$[[]]$ (да)

$] [] [$ (нет)

$[[][]]$ (да)

$[[][]][[]]$ (нет)

Пример: квадратные скобки

- Алфавит $\Sigma = \{[,]\}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

[] [] (да)

[[]] (да)

] [] [(нет)

[[][]] (да)

[[]][[]] (нет)

[[][]][[]]

Пример: квадратные скобки

- Алфавит $\Sigma = \{[,]\}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

[] [] (да)

[[]] (да)

][][] (нет)

[[][]] (да)

[[][]][[]] (нет)

[[][]][[]] [] (да)

Пример: квадратные скобки

- Алфавит $\Sigma = \{[,]\}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

[] [] (да)

[[]] (да)

][][] (нет)

[[][]] (да)

[[][][]] (нет)

[[][][]] [] (да)

€

Пример: квадратные скобки

- Алфавит $\Sigma = \{[,]\}$
- L – язык над Σ , для которого верно следующее:
 - (1) $[] \in L$;
 - (2) Для любой строки s : если $s \in L$, то $[s] \in L$;
 - (3) Для любых строк s и t : если $s \in L$ и $t \in L$, то $st \in L$
- Входят ли эти строки в L ?

$[] []$ (да)

$[[]]$ (да)

$] [] [$ (нет)

$[[[[]]]]$ (да)

$[[]] [[]]]$ (нет)

$[[]] [[]] []]$ (да)

ϵ (нет)

Важные вопросы

- Какие бывают формальные языки?
- Какие закономерности можно описать с их помощью?

- Какие бывают формальные языки?
- Какие закономерности можно описать с их помощью?

...как лингвистов, нас интересует следующее:

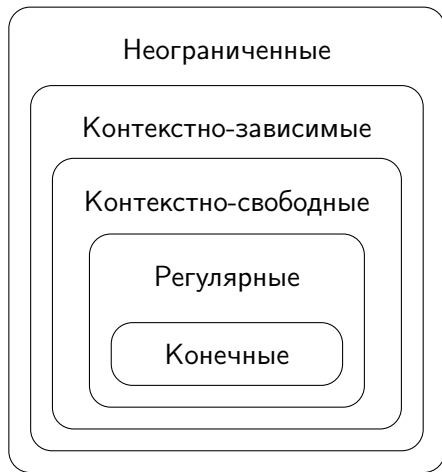
- Какие бывают формальные языки?
- Какие закономерности можно описать с их помощью?

...как лингвистов, нас интересует следующее:

- Какие формальные языки подходят для описания естественного языка?

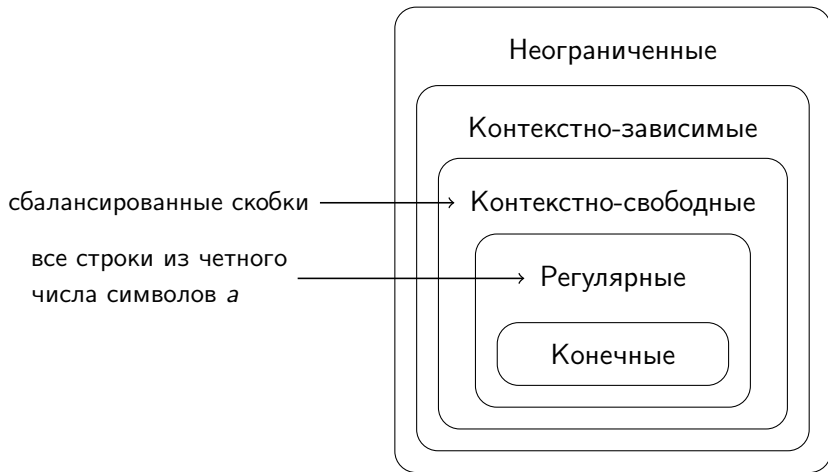
Иерархия формальных языков Хомского

(также известна как иерархия Хомского-Шютценберже)



Иерархия формальных языков Хомского

(также известна как иерархия Хомского-Шютценберже)



Чем этот курс **не** является

Чем этот курс **не** является

- Введением в лингвистику

Чем этот курс **не** является

- Введением в лингвистику
(но мы посмотрим, как можно формализовать некоторые базовые понятия лингвистики)

Чем этот курс **не** является

- Введением в лингвистику
(но мы посмотрим, как можно формализовать некоторые базовые понятия лингвистики)
- Введением в теорию формальных языков

Чем этот курс **не** является

- Введением в лингвистику
(но мы посмотрим, как можно формализовать некоторые базовые понятия лингвистики)
- Введением в теорию формальных языков
(но мы будем использовать формальные грамматики для описания закономерностей в естественном языке)

Чем этот курс **не** является

- Введением в лингвистику
(но мы посмотрим, как можно формализовать некоторые базовые понятия лингвистики)
- Введением в теорию формальных языков
(но мы будем использовать формальные грамматики для описания закономерностей в естественном языке)
- Введением в автоматическую обработку естественного языка

Чем этот курс **не** является

- Введением в лингвистику
(но мы посмотрим, как можно формализовать некоторые базовые понятия лингвистики)
- Введением в теорию формальных языков
(но мы будем использовать формальные грамматики для описания закономерностей в естественном языке)
- Введением в автоматическую обработку естественного языка
(но мы будем пользоваться программными инструментами для работы с этими грамматиками)

О чем этот курс на самом деле

- Регулярные выражения
- Конечные автоматы
- Контекстно-свободные языки
- Субрегулярные языки
- Мягко контекстно-зависимые языки
- Международный фонетический алфавит
- Фонемы и аллофоны
- Фонологические правила
- Теория оптимальности
- Автосегментная фонология
- Синтаксические составляющие
- Минималистский синтаксис

О чем этот курс на самом деле

- Регулярные выражения
- Конечные автоматы
- Контекстно-свободные языки (формальные языки)
- Субрегулярные языки
- Мягко контекстно-зависимые языки

-
- Международный фонетический алфавит
 - Фонемы и аллофоны
 - Фонологические правила
 - Теория оптимальности (лингвистика)
 - Автосегментная фонология
 - Синтаксические составляющие
 - Минималистский синтаксис

О чем этот курс на самом деле

Как связаны эти две группы терминов:

О чем этот курс на самом деле

Как связаны эти две группы терминов:

- Регулярные выражения
- Конечные автоматы
- Контекстно-свободные языки
- Субрегулярные языки
- Мягко контекстно-зависимые языки
- ...

О чем этот курс на самом деле

Как связаны эти две группы терминов:

- Регулярные выражения
- Конечные автоматы
- Контекстно-свободные языки
- Субрегулярные языки
- Мягко контекстно-зависимые языки
- ...
- Международный фонетический алфавит
- Фонемы и аллофоны
- Фонологические правила
- Теория оптимальности
- Автосегментная фонология
- Синтаксические составляющие
- Минималистский синтаксис
- ...

О чем этот курс на самом деле

Как связаны эти две группы терминов:

- Регулярные выражения
- Конечные автоматы
- Контекстно-свободные языки
- Субрегулярные языки
- Мягко контекстно-зависимые языки
- ...
- Международный фонетический алфавит
- Фонемы и аллофоны
- Фонологические правила
- Теория оптимальности
- Автосегментная фонология
- Синтаксические составляющие
- Минималистский синтаксис
- ...

...и как использовать эту связь для описания явлений фонологии, морфологии и синтаксиса естественного языка

В следующий раз...

- Формальные грамматики
- Регулярные языки (и для чего они лингвистам)