

Кому: С.Ю. Толдова, Е.А. Кузьменко

---

Копия: Компьютерные лингвисты, 3 курс

---

Тема: **Токенизация электронных писем**

---

От: Бибаева Мария, Картозия Инга, Мельник Анастасия

---



- Корпус из электронных писем курсовой почты  
(linghse2014@gmail.com)
- Объем корпуса: 3096 писем, 195307 словоупотреблений
- Языки: русский, английский.

## Тема: Output

---

Token 149: 'напримєр'

Type: cyr

Document: letter7

Token 184: 'obubnenkova@hse.ru'

Type: mail

Document: letter7

Token 156075: '#тыжлингвисте'

Type: cyr hashtag

Document: letter2565

## Тема: Что мы считаем за токен?

---

- слова ЕЯ
- числа
- буквенно-цифровые комплексы: 12-ти
- даты и время: множество форматов
- номера телефонов
- адреса: Старая Басманная, ул.
- почтовые адреса
- ссылки
- подробности в файле TZ.md

- Инициалы: А.Б., МА, Н. Р.
- Адреса: ул. Старая Басманная, 21/4, каб. 304
- Дефисы: как-то, по-русски-то, красно-зеленый
- Номера телефонов: 8(495) 772-95-90\*222-80

## Тема: Тестирование и output

---

Точность: 0.99564

Полнота: 0.99132

Token 1: 'Здравствуйте'

Type: cyr lex-prop

Document: letter 800

Token 2: '!'

Type: punctuation

Document: letter 800

Token 3: 'В'

Type: cyr

Document: letter 800