

Автоматизация построения
языковых моделей для
распознавания спонтанной
русской речи



Белик В. В., Бирин Д. А.
ФГУП «НИИ «КВАНТ», Санкт-Петербург.

Этапы построения ЯМ



▼	□	Нормализация текстов корпуса (СНТ) ▾	Run ...			
#399	🔄	Updating sources ▾	No artifacts ▾	Белик Виктория (1) ▾	45m:20s left	Stop
#398	✅	Success ▾	Artifacts ▾	No changes ▾	5 days ago (12m:34s)	
▼	□	Разделение корпуса на обучающую и тестовую часть ▾	Pending (13) ▾	1 queued ▾	Run ...	
#166	✅	Success ▾	No artifacts ▾	Changes (4) ▾	one month ago (5h:02m)	
▼	■	Построение частотных словарей ▾	Run ...			
#301	🔄	Step 3/5 ▾	No artifacts ▾	Белик Виктория (1) ▾	2s left	Stop
#300	❗	Exit code 1 (new) ▾	No artifacts ▾	No changes ▾	6 days ago (36s)	
▼	□	Подсчёт n-грамм ▾	Pending (1) ▾	Run ...		
#509	✅	Success ▾	Artifacts ▾	No changes ▾	5 days ago (42m:25s)	
▼	□	Построение ЯМ ▾	Run ...			
#503	✅	Success ▾	No artifacts ▾	No changes ▾	4 days ago (15s)	
▼	□	Оценка качества ЯМ ▾	Run ...			
#670	✅	Tests passed: 3 ▾	Artifacts ▾	No changes ▾	2 days ago (3h:36m)	

Операции предобработки



- ↻ замена случайных вставок букв латинского алфавита на кириллические символы;
- ↻ исправление регистра;
- ↻ ёфикация;
- ↻ удаление недопустимых символов и комбинаций символов;
- ↻ отделение знаков препинания пробелами;
- ↻ разделение текста на предложения, каждое из которых занимает отдельную строку;
- ↻ исправление ошибок и опечаток.

Пример предобработки

До предобработки

"> **Происшествия**
Россия и Турция обсуждают возможные варианты диалога в Астане по Сирии
 28 декабря 2016, 13:27
 Москва и Анкара постоянно контактируют и обсуждают возможные темы переговоров по Сирии, которые планируется провести в середине января в Астане. Об этом сообщил пресс-секретарь президента России Дмитрий Песков. Действительно, ведутся постоянные контакты с турецкими коллегами, обсуждаются различные модальности возможного диалога, который планируется в Астане. Все это ведется в русле поиска политического урегулирования в Сирии , — сказал Дмитрий Песков <a class="tags-t

После предобработки

<s> происшествия . </s>
<s> Россия и Турция обсуждают возможные варианты диалога в Астане по Сирии . </s>
<s> Москва и Анкара постоянно контактируют и обсуждают возможные темы переговоров по Сирии которые планируется провести в середине января в Астане . </s>
<s> об этом сообщил пресс-секретарь президента России Дмитрий Песков . </s>
<s> действительно ведутся постоянные контакты с турецкими коллегами обсуждаются различные модальности возможного диалога который планируется в Астане . </s>
<s> всё это ведётся в русле поиска политического урегулирования в Сирии сказал Песков . </s>

Созданные корпуса

Новостной



Разговорный

- ☞ 4 555 454 документа
- ☞ 34 983 030 предложений
- ☞ 1 219 703 306 словоформ
- ☞ 12 803 источника

- ☞ 8 932 домена
- ☞ 16 Гб

- ☞ 8 970 документов
- ☞ 5 631 829 предложений
- ☞ 36 447 342 словоформ
- ☞ 8 источников

- ☞ 2 вспомогательных корпуса:
Субтитры 15,3 млн словоформ
Форумы 4,3 млн словоформ

Полученные модели

Новостная

Разговорная



710 977

■ **Объем словаря**

236 378

2+

■ **Отсечение словаря**

Без отсечения

89 267 067

■ **Количество биграмм**

8 165 407

104 564 900

■ **Количество триграмм**

5 431 113

2+

■ **Отсечение триграмм**

2+

Результаты тестирования ЯМ



Новостная

Разговорная

0

■ Медиана пропуска

2

36,87

■ Медиана связности

294

25

■ Покрытие биграммami

47

66

■ Покрытие триграммами

39