



Автоматическая обработка текста 1.

# Подготовка корпуса Препроцессинг

- Корпус:
  - Learning set
  - Development set
  - Testing set
  - (NB метод кроссвалидации)
- Разметка корпуса:
  - Инструкция
  - Мера согласия между аннотаторами
  - Золотой стандарт
  - Разметка обучающего множества / тестового множества

## 1. Препроцессинг

- графическая нормализация
- токенизация
- сегментация на предложения

## 2. Дополнительная обработка

- индекс
- оффсеты
- классификация токенов

## 3. Распознавание языка

## 1. Препроцессинг

- графическая нормализация
- токенизация
- сегментация на предложения

## 2. Дополнительная обработка

- индекс
- оффсеты
- классификация токенов

## 3. (Распознавание языков)

## Пример 1.

**Л.** Қалтаева жұмыспен қамту бағдарламасын «еңбек етіп табыс тауып, өзін лайықты, сенімді адам сезінгісі келетін» мүгедектігі бар адамдар үшін «жақсы трамплин» болып табылады деп санайды екен. **<br>**– 2013 жылғы 1 қаңтардағы жағдай бойынша Қазақстанда барлығы 609 мыңнан астам мүгедек бар, – деді Денсаулық сақтау министрі Салидат Қайырбекова. **<br>**18 жасқа дейінгі мүгедек балалар саны – 65 мың 844, оның ішінде 57 мың 627-сі – 16 жасқа дейінгі балалар.

## Прмер 2.

Пример: *<i>niceweather</i> =&gt; <i>nice weather</i><br>* Но, для *<i>workinggrass</i>* сегментация опять же не будет найдена. Первое слово, которое заматчит наш алгоритм будет *<i>working</i>*, а не *<i>work</i>* и которое также поглотит первую букву в слове *<i>grass</i>*.**<br>**

Может нужно скомбинировать каким-то образом оба алгоритма? Но, как тогда быть со строкой *<i> niceweatherwhenworkinggrass >*? В общем пришли к брутфорсу.**<br>**

## Пример 3.

Таскаю в Воспитательный  
Своих незаконнорожденных детей...

<"ВИД ИМЕНИЯ ГУРЗУФ...">

{ Рис. 1 }

АТТЕСТАТ

Рукой А. Янова:

Сего февраля 24 дня осмотрена мною подушка черного атласа размером приблизительно 1/2 аршина в квадрате. Подушка с вышивкой по наложенным кретоновым цветам, среди коих помещается овал, также кретоновый, вырезан в овале сюжет буколического содержания в две фигуры (не считая собаки\*): пастушка, разговаривающая с пастушком, на дальнопланной части кусты и деревья.

Все сие в общем кретоне вышивка шелком, черный фон атласа, красный шнур по швам подобраны с отменным вкусом и тщательностью. Особенно удачно вышли цветы. Вообще все хорошо, о чем свидетельствую своим подписом. Да! вообще все хорошо!!..... Да, хорошо...

\* или, может, козочки

## Пример 4.

[risingvoices](#) [@risingvoices](#)

Glad that our friends -)) (and partners for events like Tweet [#MotherLanguage](#)) from [@IndigenousTweet](#) & [@livingtongues](#) will be at [#OxfordGL](#)

Retweeted by [Indigenous Tweets](#)

# Предварительная обработка текста

1. Удаление нетекстовых элементов (остатки HTML и других служебных «не текстовых» символов)
2. Исправление стандартных ошибок распознавания: кириллица vs. латиница
3. Стандартизация символов: тире, кавычки, пробелы
4. Артефакты конвертации в другой формат
5. Выделение и оформление нестандартных (нелексических) элементов, например:
  - элементов форматирования – жирность, курсивность, подчёркивание;
  - структурных элементов текста – заголовков, абзацев, примечаний;
  - различных элементов текста, не являющихся словами (числа, даты в цифровых форматах, буквенно-цифровые комплексы, и т.п.);
  - имен (имя, отчество), написанных инициалами;
  - иностранных лексем, записанных латиницей и т.д.
6. Сборка (например, слов, написанных в разрядку)





1. Распознавание языка
2. Препроцессинг
  - графическая нормализация
  - **токенизация**
  - сегментация на предложения
3. Дополнительная обработка
  - индекс
  - оффсеты
  - классификация токенов
4. (Исправление ошибок)

## Графематический анализ

- Уровень обработки: символы текста
- Графема – единица текста (письменного), неделимый знак (буквы, знаки препинания и др.)
- Цель – выделение и классификация основных единиц текста: слов, предложений, абзацев

- **Токены** – единицы обработки соответствующие словам («псевдословам»)
- Token – кусочек, обычно цепочка знаков от пробела до пробела (компиляция ЯП: лексема)
- Цель – выделение минимальных лингвистически значимых элементов текста (токенов)
- Виды токенов:
  - слова ЕЯ
  - знаки препинания
  - обозначения денежных единиц
  - числа
  - буквенно-цифровые комплексы
  - даты (множество форматов)
  - номера телефонов
  - IP -адреса и имена файлов

- Наивная токенизация – все разбиваем по пробелам
- Всегда ли пробелы имеют одну и ту же функцию:
  - ✓ Both “Los Angeles” vs. “rock 'n' roll”
  - ✓ 100 000
- !!!!
- Обычно считается, что токенизация – очень простая, неинтересная и не очень значимая процедура
- НО: от качества токенизации может очень сильно зависеть качество высокоуровневых задач
- Ср. выделение именованных сущностей: U.S., ex-president, don't

- Основания для уточнения правил токенизации:
  - а) языковая реальность
  - б) конечная NLP задача
  - с) архитектура обработки
- а) критерий слова -))),
  - ср. русск. *Петя-то пришел*  
*Пойди-ка принеси*
  - ср. *Time-frame, timeframe, time frame*
  - Интерпретация токена может зависеть от контекста:
    - 100 г. / г. Ростов - dr – doctor / drive

- регистр (обычно все приводим к одному регистру)
- знаки препинания и служебные символы
- обработка точки
- обработка дефиса
- обработка апострофа
- обработка буквенно-числовых комплексов
- обработка дат – нормализация дат
- **ТИПИЗАЦИЯ ТОКЕНОВ**
- **офсеты**

- аббревиатуры
- ГОТОВЫЕ СПИСКИ И СЛОВАРИ АКРОНИМОВ (точка элемент сокращения)
- ТОЧКА - ОТДЕЛЬНЫЙ ТОКЕН
- ОМОНИМИЯ:
- `in' - `inches; `no' - `number, `bus' - `business; `sun' - `Sunday;
- ПО

# Токенизация: аббревиатуры

- O.U.
- M.D.
- N.B.
- P.O.
- U.K.
- U.S.
- U.S.A.
- P.S.
- mr.
- mrs.
- .com
- dr.
- .sh
- .java
- st.
- .net

<https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en>



## Обработка дефисов

- self-assessment,
- F-15,
- forty-two
- Los Angeles-based.
- !!! Зависит от задачи
- Частеречные разметчики (part-of-speech taggers)
- NER – скорее разные слова: *'Moscow-based'*

<https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en>

- Типы дефисов

- переносы слов

- лексические

- Элементы словаря *Тянь-шань*

- Стандартное написание некоторых префиксов *co-, pre-, meta-, multi-, etc.*

- Модели: *forty-seven*

- Окказионализмы: *end-of-line*

- “Синтаксические”

- `ed'-отглагольные прилагательные *case-based, computer-linked, hand-delivered*

- *three-to-five-year*

- *the New York-based co-operative was fine-tuning forty-two K-9-like models.*

Token	Type
New York-based	<b>Sentential</b>
co-operative	<b>Lexical</b>
fine-tuning	<b>End-of-Line</b> , but could also be considered a <b>Lexical</b> hyphen based on the author's stylistic preferences.
Forty-two	<b>Lexical</b>
K-9-like	<b>Lexical and Sentential</b>

# Токенизация: обработка буквенно-цифровых комплексов

- Examples: Email addresses
- URLs
- Complex enumeration of items
- Telephone Numbers
- Dates
- Time
- Measures
- Vehicle Licence Numbers
- Paper and book citations
- etc

## Телефонные номера

- 123-456-7890
- (123)-456-7890
- 123.456.7890
- (123) 456-7890
- etc

<https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en>

- Отдельные модули для распознавания и нормализации таких классов объектов (телефонных номеров, интернет адресов, дат)
- Date/Time Formats:
  - 8th-Feb
  - 8-Feb-2013
  - 02/08/13
  - February 8th, 2013
  - Feb 8th
  - И т.п.
- <https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en>
- см., например, <https://www.kaggle.com/c/text-normalization-challenge-russian-language>

# Токенизация: стандарты

*"I said, 'what're you? Crazy?'" said Sandowsky. "I can't afford to do that."*

<https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en>

# Tokenization Example

	по пробелам	Apache Open NLP (en-token.bin)	Stanford 2.0.3	Custom	Гипотетический (3C)
1		“	“	“	“
2	“I	i	i	i	i
3	said,	said	said	said	said
4		,	,	,	,
5	'what're	'what	`	'	'
6			what	what're	what
7		're	're		are
8	you?	You	you	You	you
9		?	?	?	?
10	crazy?"	crazy	crazy	crazy	crazy
11		?	?	?	?
12		'	'	'	'

	Naïve по пробелам	Apache Open NLP	Stanford 2.0.3	Custom	Гипотетич. (3C)
13	Said	said	Said	said	said
14	sandowsky.	sandowsky	sandowsky	sandowsky	sandowsky
15		.	.	.	.
16		,	,	,	“
17	'i	i	i	i	i
18	can't	ca	ca	can't	can
19		n't	n't		not
20	afford	afford	afford	afford	afford
21	to	to	to	to	to
22	do	do	do	do	do
23	that.'	that	that	that	that
24		.	.	.	.
25		,	,	,	,

<https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en>

# Токенизация: пунктуация внутри токенов

- u.s.a.,
- Ph.D.,
- AT&T,
- ma'am,
- cap'n,
- 01/02/06
- stanford.edu
- 7.1
- "\$2,023.74"

<https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en>



# Токенизация: задачи и архитектура системы

- Named Entity Extraction
- `<node span="Rational Software Architect for WebSphere" type="NP">  
 <node span="Rational Software Architect for WebSphere"  
 type="NNP"/>  
 </node>`

The abbreviated forms of *be*:

1. 'm in I'm
2. 're in you're
3. 's in she's

The abbreviated forms of *auxiliary verbs*:

1. 'll in they'll
2. 've in they've
3. 'd in you'd

Note that clitics in English are ambiguous. The word "she's" can mean "she has" or "she is".

A tokenizer can also be used to expand clitic contractions that are marked by apostrophes, for example:

what're => what are

we're => we are

<https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en>

- Шаг 1. Разбиение по пробелам, очистка от кавычек, скобок и др. служебных СИМВОЛОВ
- Шаг 2.
- Обработка сокращений и точек в сокращениях (в некоторых приложениях точка сохраняется как значимый символ аббревиатуры)
- Шаг 3.
- Дефисы
- Шаг 4.
- Обработка бувенно-числовых и числовых комплексов
- Шаг 5.
- Обработка дат



- Языки с пробелами
  - Специальные случаи для обработки
- Беспробельные языки



- #habratopic => habra topic
- **geschwindigkeitsbegrenzung** — *ограничение скорости*
- **城市人的心爱宠物** — *любимое домашнее животное городских жителей*

## Алгоритм 1. Minimum Matching

- *niceday*
- *niceweather*

Проходимся по строке и находим первое слово, которое совпадает со словарным. Сохраняем это слово и повторяем процедуру для остатка строки. Если в последней строке не находится ни одно совпадающее со словарным слово, считаем что сегментация не найдена.

*nice + we + a + the + ??? - X*

<https://habrahabr.ru/post/141228/>

## Алгоритм 2. Maximum Matching или Greedy

- *niceweather*
- *Workingrass*

Вначале выбираем максимально длинное слово

Идем с конца строки.

(медленнее, чем Minimum matching)

- *working + grass - \**
- <https://habrahabr.ru/post/141228/>

- Алгоритм 3. Все варианты разбиения по словарю
- *expertsexchange* => (*expert sex change*, *experts exchange*)
- *dwarfstealorcore* -
  - «дворф крадет или ядро»?
  - «дворф крадет руду орков»

+ Униграмная модель – частоты употребления каждого токена (слова) в тексте  
*input* => *in put*

[https://github.com/swigder/word\\_segmentation](https://github.com/swigder/word_segmentation)





# Типизация токенов

## Пример. Описание токена в среде OntosMiner

извлечённых из электронных толковых словарей // URL:

Token		
kind	word	X
lang	cyr	X
length	11	X
orth	lowercase	X
string	извлечённых	X
		X

Open Search & Annotate tool

О.Н., Толдова С.Ю. Снятие семантической многозн: // URL:

Token		
kind	word	X
lang	cyr	X
length	1	X
orth	upperInitial	X
string	О	X
		X

Open Search & Annotate tool

DocEnd

Token		
kind	word	X
lang	lat	X
length	3	X
orth	allCaps	X
string	URL	X
		X

Open Search & Annotate tool

Paragraph



- Основные графематические дескрипторы (19):
- OLE – русская лексема (последовательность букв кириллицы)
- OILE – иностранная лексема;
- OPun – знак препинания (. , : ; -);
- OOpn и OCls – открывающая и закрывающая скобки; Контекстные дескрипторы (19)
- OSent1 и OSent2 – признаки начала и конца предложения;
- OBulet – признак начала пункта перечисления;
- OPar – признак начала абзаца;
- OFIO1 и OFIO2 – признак начала и конца ФИО;
- Дескрипторы макросинтаксического анализа (5):
- CS\_Heading – признак конца заголовка;
- CS\_Parent – конца раздела, заканчивающегося знаком :

# Токенизация: адреса токенов

Sentence № 12835  Arrows  Tables

Милиционер поднял голову , увидел пуле

<i>SyntAutom</i>			
id	token	type	head
1	<u>Милиционер</u> ← <i>поднял</i>	subj:nom	2
2	<u>поднял</u>	root	
3	<u>голову</u> ← <i>поднял</i>	obj:acc	2
4	, ← <i>поднял</i>	conj	2
5	<u>увидел</u> ← <i>поднял</i>	coord	2
6	<u>пулевое</u> ← <i>отверстие</i>	adj	7

# Токенизация: адреса токенов

Cancel

Var: пе

Gr: 9:1083[8]

« ШКОЛЫ ... »

Head: Школы

Link: 8:819[5]

properties

ref

дальше, тем больше «родства» наблюдалось у ведущих «Школы...».

10: Потом проект переехал с «Культуры» на НТВ.

11: Это помимо явных перемен в виде тут же появившихся рекламных

тон разговора.

Матвей Курзуков      ✗ Саша Печеный      ✗

Key	Group	ref	str	type	Ref	Group	ref	str	type
8:788[17]	« Школа злословия »	def	noun	coref	3:13[17]	« Школа злословия »	def	noun	coref
8:819[5]	собой	def	refl	coref	4:96[9]	программы	def	noun	coref
9:1084[5]	Школы	def	noun	coref	5:187[8]	« Школа ... »	def	noun	coref
10:1099[6]	проект	def	noun	coref	7:734[9]	программы	def	noun	coref
12:1335[9]	программы	def	noun	coref	8:788[17]	« Школа злословия »	def	noun	coref
13:1441[17]	« Школы злословия »	def	noun	coref	8:819[5]	собой	def	refl	coref
17:2064[17]	« Школа злословия »	def	noun	coref	9:1083[8]	« Школы ... »	def	noun	coref

- Шаг 1. Разбиение по пробелам, очистка от кавычек, скобок и др. служебных СИМВОЛОВ
- Шаг 2. Обработка сокращений и точек в сокращениях (в некоторых приложениях точка сохраняется как значимый символ аббревиатуры)
- Шаг 3. Дефисы
- Шаг 4. Обработка бувенно-числовых и числовых КОМПЛЕКСОВ
- Шаг 5. Обработка дат
- Типицация токенов
- Адреса токенов

# Форматы представления текстовых корпусов

```
<document>  
  <docID>040404-27793</docID>  
  <docURL> URL документа в Веб в base 64</docURL>  
  <subject encoding="base64"> тема новости в base64 </subject>  
  <agency>название новостного агентства в base64</agency>  
  <timestamp>  
    <date>20040402</date>  
    <daytime>50493</daytime>  
  </timestamp>  
  <content encoding="base64">  
    содержимое в base64  
  </content>  
</document>
```

# Форматы представления текстовых корпусов

Stand-off разметка: файл с текстами, как они есть, файл с фрагментами, на которые нужно положить аннотацию, адресами и признаками

```
<resultlist>
```

```
<entitylist srcRef="1-27793">
```

```
<entity class="person" offset="432" length="18" id="1">Эдуарда Шеварднадзе</entity>
```

```
<entity class="other" offset="198" length="13" id="2">революция роз</entity>
```

```
</entitylist>
```

```
<entitylist srcRef="2-45913">
```

```
<entity class="organization" offset="1" length="19" id="1">большая восьмерка</entity>
```

```
<entity class="place-name" offset="56" length="9" id="2">Шотландия</entity>
```

```
<entity class="organization" offset="320" length="3">G8</entity>
```

```
<entity class="organization" offset="548" id="1">большой восьмерке</entity>
```

```
</entitylist>
```

```
</resultlist>
```

# Форматы представления текстовых корпусов. CoNLL

## CoNLL

# sent\_id = dev-143

# text = Jeg siger til ham , at min længsel efter ham er stærkere end smerten .

```

1  Jeg  jeg  PRON  _      Case=Nom|Gender=Com|Number=Sing|Person=1|PronType=Prs  2  nsubj
2  siger  sig  VERB  _      Mood=Ind|Tense=Pres|VerbForm=Fin|Voice=Act  0  root
3  til  til  ADP  _      AdpType=Prep  4  case  _  _
4  ham  han  PRON  _      Case=Acc|Gender=Com|Number=Sing|Person=3|PronType=Prs  2  obl
5  ,  ,  PUNCT  _  _      2  punct
6  at  at  SCONJ  _  _      12  mark
7  min  min  DET  _
   Gender=Com|Number=Sing|Number[psor]=Sing|Person=1|Poss=Yes|PronType=Prs  8  det
8  længsel  længsel  NOUN  _      Definite=Ind|Gender=Com|Number=Sing  12  nsubj
9  efter  efter  ADP  _      AdpType=Prep  10  case
10 ham  han  PRON  _      Case=Acc|Gender=Com|Number=Sing|Person=3|PronType=Prs  8  nmod

```

<http://universaldependencies.org/docs/format.html>



# Форматы представления текстовых корпусов. CoNLL

ID FORM LEMMA PLEMMA POS PPOS FEAT PFEAT HEAD PHEAD DEPREL PDEPREL

ID (index in sentence, starting at 1)

FORM (word form itself)

LEMMA (word's lemma or stem)

POS (part of speech)

FEAT (list of morphological features separated by |)

HEAD (index of syntactic parent, 0 for ROOT)

DEPREL (syntactic relationship between HEAD and this word)

<http://universaldependencies.org/docs/format.html>

Используются:

- Маркеры конца предложения – точка, вопросительный и восклицательный знаки но! – неоднозначны (сокращения слов, инициалы, сокращения в конце предложения: Dr.White )
- Маркер начала предложения – заглавная буква, также неоднозначен (цитаты и др.) ..пр. Мишка прыгал по полу... сказал он: «Я вижу лес..
- Цитаты (прямая речь), оформляются в разных языках кавычками /апострофами (в англ. языке - также для обозначения притяж. падежа и сокращений : Ann's , it's )
- Требуется анализ локального контекста маркеров
- Точность сегментации зависит от тематики и жанра текстов, количества сокращений, имен собственных
- Применение машинного обучения (статистика по корпусам)

## Машинное обучение

*(1) Но ведь Маша знает А. Б. Иванова много лет и никогда про него ничего плохого не слышала!!*

- сам знак препинания (punct),
- ближайшее псевдослово слева (left),
- ближайшее псевдослово справа (right)
- ближайшее собственно слово справа (wright).

Под псевдословом здесь и далее понимается любая последовательность символов, не включающая пробел или конец абзаца.

Под словом – псевдослово, содержащее хотя бы одну букву или цифру. Также запоминается количество псевдослов слева (dleft) и справа (dright) до ближайшей потенциальной границы или конца абзаца.

## Машинное обучение

*(1) Но ведь Маша знает А. Б. Иванова много лет и никогда про него ничего плохого не слышала!!*

punct	.	.	!	!
left	А	Б	слышала	!
right	Б	Иванова	!	-отсутствует-
wright	Б	Иванова	-отсутствует-	-отсутствует-
dleft	4	1	11	0
dright	1	11	0	0

- сам знак препинания (punct),
- ближайшее псевдослово слева (left),
- ближайшее псевдослово справа (right)
- ближайшее собственно слово справа (wright).

## Машинное обучение

*(1) Но ведь Маша знает А. Б. Иванова много лет и никогда про него ничего плохого не слышала!!*

- словарь сокращений  
(для этого из размеченной части Национального Корпуса Русского Языка (около 6 миллионов слов) были извлечены все триграммы вида «псевдослово точка слово\_со\_строчной\_буквы»)
- каждый контекст проверяется на сокращения: если псевдослова `left` или `right` нашлись в словаре, то контекст получает дополнительные признаки `abbleft` и `abbright` соответственно
- классификация псевдослов (пунктуация, числа; остальные псевдослова были разбиты на классы в зависимости от используемых символов: кириллица, латиница, кириллица+латиница, кириллица+цифры и так далее и регистра первого символа (строчная буква, прописная буква, цифра, пунктуация))
- класс псевдослов описывается признаками `cleft` и `cright`
- ближайшее собственно слово справа, `wright`, описывается одним дополнительным признаком – регистр первого символа (`cwright`)
- дополнительные признаки `isfirst` и `islast` для обозначения контекстов, приходящихся на начало или конец абзаца.

## Сегментация на предложения

punct	.	.	!	!
left	А	Б	слышала	!
right	Б	Иванова	!	-отсутствует-
wright	Б	Иванова	-отсутствует-	-отсутствует-
dleft	4	1	11	0
dright	1	11	0	0
ableft	1	1	0	0
abbright	1	0	0	0
cleft	-кириллица- -прописная-	-кириллица- -прописная-	-кириллица- -строчная-	
cright	-кириллица- -прописная-	-кириллица- -прописная-	-пунктуация-	-отсутствует-
cwright	-прописная-	-прописная-	-отсутствует-	-отсутствует-
isfirst	0	0	0	0
islast	0	0	0	1

	Эксперимент 1		Эксперимент 2	
	точность, %	полнота, %	точность, %	полнота, %
termpunct	67.2	100**	66.9	98.9**
termpunct_cap	90.7	97.0**	89.6	96.0**
advanced	96.4	90.4	95.0	89.6
C4.5	97.8	98.5**	98.5*	97.5**
Ripper	98.5	98.5**	98.9**	96.0**
SVM-light	99.6**	98.5**	99.6**	97.5**

- termpunct - только контексты, содержащие терминальные знаки препинания.
- termpunct\_cap - запрещает предложения, не начинающиеся с заглавной буквы.
- advanced - запрещает предложения, заканчивающиеся сокращением и точкой

- Таким образом, при сегментации нужны компоненты:
  - Словарь сокращений
  - Словарики графических знаков
  - Словарь устойчивых оборотов (обычно более 500)
  - Эвристические правила анализа контекстов, более чем один просмотр текста
  - Языково-зависимые компоненты !
  - (также зависит от тематики текстов, причины - различная роль знаков препинания и др.)
  - Достигаемая точность – до 99, полнота – 60-80 %

Восточные языки ( non-segmented languages) - слитное написание слов

⇒ применяются:

- статистические методы сегментации, морфословари, грамматические правила (японский) , также европ. языки с большим сложносоставных слов, например, немецкий: Worterbuch



# Технологии реализации графематического анализа

- Формальный аппарат на базе теории формальных языков и грамматик
- Простейший графематический анализ – анализ регулярных языков (Тип 3 по Хомскому)
- Более сложный граф. анализ – учет локального контекста, словари
- Средства описания регулярных языков
  - Регулярные выражения
  - Регулярные (автоматные) грамматики
  - Конечные автоматы

- Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.
- Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
- Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.
- Oxford Handbook on Computational Linguistics. R. Mitkov (Ed.). Oxford University Press, 2005 , p. 201-218.
- <http://sentiment.christopherpotts.net/tokenizing/>