# 8 тема:

- 1) RuCor (http://rucoref.maimbava.net/),
- 2) AnCora (http://clic.ub.edu/corpus/),
- 3) ARRAU (https://catalog.ldc.upenn.edu/LDC2013T22),
- 4) Мультимедийный корпус идиш (http://web-corpora.net/YiddishMultimediaCorpus/search/),
- 5) Транслитератор идиш (http://web-corpora.net/YiddishTransliterator/)

### 1) RuCor (http://rucoref.maimbava.net/)



### RuCor

Russian coreference corpus

#### Links

(i) rucoref.maimbava.net

Corpus download, version of 2015-10-29 (3.8mb) Web interface

### Corpus description

RuCor is the first open corpus of Russian language where anaphorical and coreferential relations between noun groups are annotated. The current version of RuCor contains 156636 tokens. Apart from the annotation of coreferential and anaphorical relations morphological annotation is also provided.

The elaboration of RuCor started in 2013 as a part of the project RUEVAL2014, campaign evaluating the quality of Russian NLP tools to resolve anaphora and extract coreference chains.

RuCor includes prosaic texts of different length and genres; news, science, fiction, blogs,

This resource is aimed at theoretical linguists working in the field of anaphora and coreference as well as at NLP systems' developers and at all those who are fascinated by Russian syntax and discourse.

All materials are open and available for download. If you quote examples retrieved from RuCor, please, cite RuCor as the source as well as the author of the text in question and the name of the text.

The Web interface was designed by Dmitrij Gorshkov. The tool uses MySQL database engine for corpus management.

#### Corpus users

Our target audience are specialists in theoretical and applied linguistics, students and lecturers in linguistics.

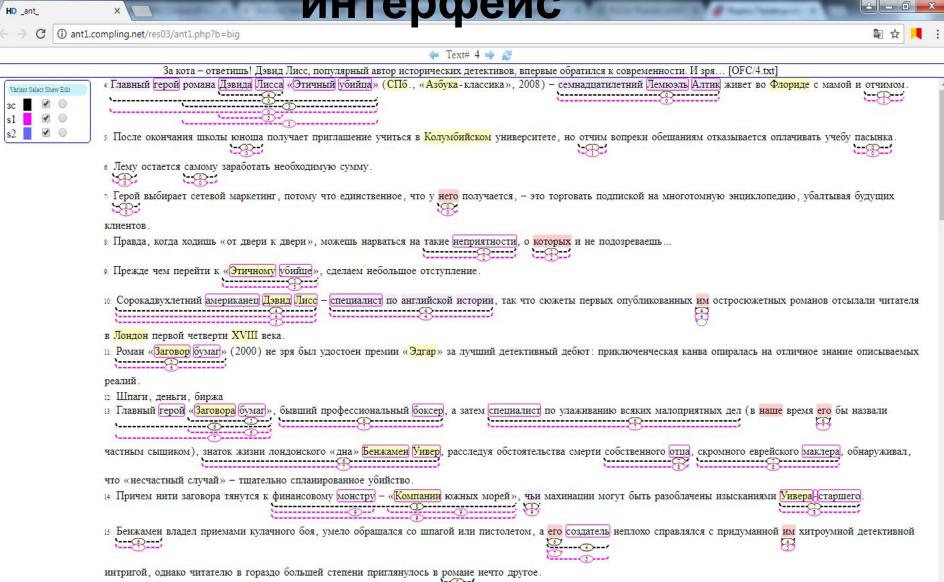
RuCor can be used for a variety of purposes in theoretical research: primarily, for narrow-oriented studies of anaphora and coreference, but also for more global studies of syntax and discourse structure, typology of anaphora, cognitive aspects of reference and referential choice.

Texts taken from this corpus can serve lecturers and students as data during seminars and lectures dedicated to corpus technologies in applied linguistics and dedicated to anaphora and coreference in discourse. Psycholinguists might be tempted by the possibility to determine factors influencing referential choice in different types of texts.

NLP developers can use texts of RuCor as a training set for machine learning algorithms of anaphora and coreference resolution or as a golden standars to evaluate the success of their software.

Корпус	1) Web формат 2)Можно скачать
Взаимосвязь в предложении	<ul> <li>- анафорические и кориферентные отношения между существительными группами аннотируются</li> <li>- также предусмотрена морфологическая аннотация</li> </ul>
Формат	RuCor включает в себя прозаические тексты различной длины и жанров: новости, наука, фантастика, блоги.
Аудитория	<ul> <li>Данный ресурс направлен на теоретиков-лингвистов, работающих в области анафоры и корреляции, а также на разработчиков систем НЛП и всех тех, кто увлекается русским синтаксисом и дискурсом.</li> <li>специалисты в области теоретической и прикладной лингвистики, студенты и преподаватели в области лингвистики</li> </ul>
RuCor использовние	первую очередь, для узко-ориентированные исследования анафоры и кореферентности, но и для более глобальных исследований синтаксиса и структуры дискурса, типологии анафоры, когнитивные аспекты ведения и ссылочной выбор.
Общая статистика	количество текстов 181 количество маркеров 156637 Количество цепей корреляции 3638 количество выбранных существительных групп 16558 Распространение текстовых жанров: <новости 45% эссе 21% фантастика 18% науки 9%блогов, комментариев 5% Русская Википедия 2%

Веб-интерфейс



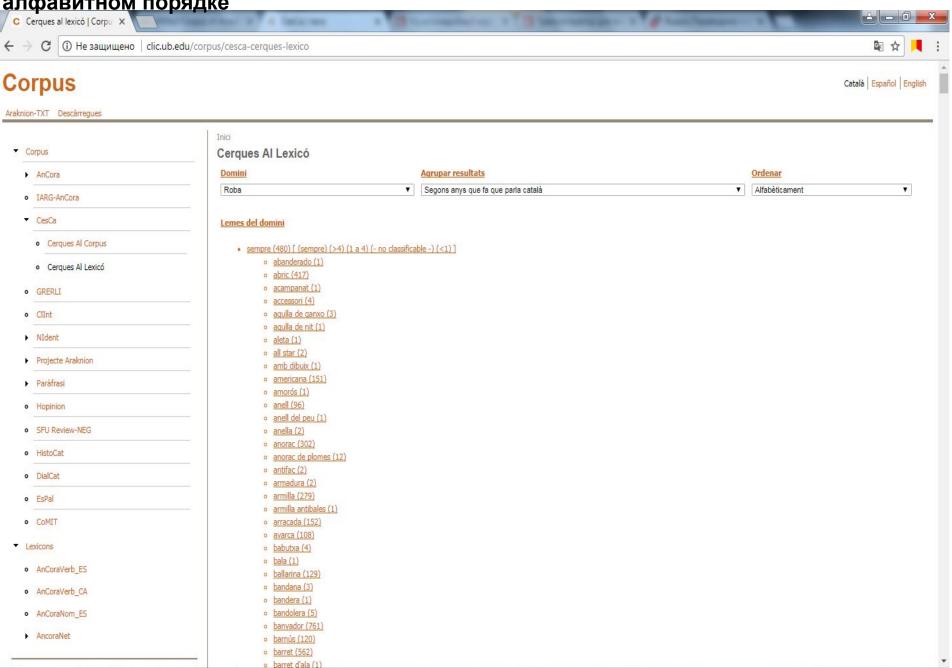
### 2) AnCora (http://clic.ub.edu/corpus/)

Ancora-это свод каталанский (анкора-СА) и испанский (анкора-ЭС) с различными

Кол-во слов	15 <i>766 265</i> предложений с <i>537 871 550</i> словами
Тексты	текстовые документы, извлеченные из источников, начиная от актов испанского парламента и заканчивая испанской версией Википедии
Жанры	энциклопедические тексты, газетные статьи, парламентские акты, реплики королевских домов, новости из пресс-агентства, книги, новости общества

- -девиз и морфологическая категория
- -составляющие и синтаксические функции
- -структура сюжета и тематические документы
- -семантический словесный класс
- -денотативный тип девербальных имен
- -Номинальное слово
- -назначенные лица
- -корреляционные соотношения

Похожие формы, употреб. с исходным словом, найдены с исходного языка в алфавитном порядке



### 3) ARRAU (https://catalog.ldc.upenn.edu/LDC2013T22)

Цель: по лингвистическим данным поддерживать языковое образование, исследования и развитие технологий путем создания и обмена лингвистическими ресурсами.

- каталог ежегодно растет на 30-36 корпусов и содержит данные

Источники:	Материал с текстом, новости, социальные сети
Возможности:	анализ, извлечение информации, обнаружение информации, анализ дискурса, теги
Язык:	английский
минусы:	<ul><li>- Английский язык</li><li>- Обязательная регистрация</li></ul>



Linguistic Data Consortium UNIVERSITY OF PENNSYLVANIA

■ Мультимедийный корп × ▼ Транслитератор для язь ×

CONTACT US











ABOUT	What's New:	
MEMBERS		
COMMUNICATIONS	LDC closed for Thanksgiving, November 23-26	
LANGUAGE RESOURCES		
DATA MANAGEMENT	Join LDC for Membership Year 2018	
COLLABORATIONS		
	Spring 2018 Data Scholarship Program	
Quick Links	Web pages feature DMDs	
CATALOG	Web pages feature DMPs	
NEW CORPORA	LDC enhances its user services	
USER LOGIN		
HOW TO GET DATA	Staff Podcasts Accessible on the LDC Blog	
DATA MANAGEMENT PLANS		
PROJECTS		

#### How LDC Data Inspires Research

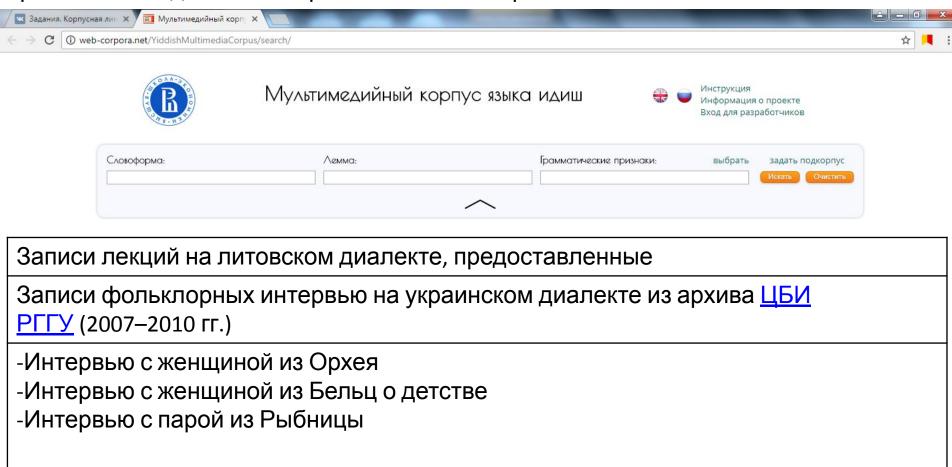
```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE nitf SYSTEM "http://www.nitf.org/IPTC/NITF/3.3/specification/dtd/nitf-3-3.dtd">
<nitf change.date="June 10, 2005" change.time="19:30" version="-//IPTC//DTD NITF 3.3//EN">
 <title>Irabu, at His Best, Silences Favorite Foe</title>
 <meta content="02YANK$03" name="slug"/>
 <meta content="2" name="publication_day_of_month"/>
 <meta content="7" name="publication_month"/>
 <meta content="1999" name="publication_year"/>
 <meta content="Friday" name="publication_day_of_week"/>
 <meta content="Sports Desk" name="dsk"/>
 <meta content="1" name="print_page_number"/>
 <meta content="D" name="print_section"/>
 <meta content="3" name="print_column"/>
 <meta content="Sports" name="online_sections"/>
 <docdata>
  <doc-id id-string="1120326"/>
  <doc.copyright holder="The New York Times" year="1999"/>
  <series series.name="BASEBALL"/>
  <identified-content>
   <classifier class="indexing_service" type="descriptor">Baseball</classifier>
   <org class="indexing_service">New York Yankees</org>
   <org class="indexing_service">Detroit Tigers</org>
   <person class="indexing_service">Olney, Buster</person>
```

(Sample from the New York Times Annotated Corpus)

The New York Times Annotated Corpus illustrates how data published in LDC's Catalog can become an important resource for the community. The New York Times is one of LDC's earliest data providers; the billions of words of news text it has provided for language resources since the 1990s continue to be used today for research and technology development. Its contribution of the New York Times Annotated Corpus in 2008 opened a new dimension for research with summaries, tags and parsing tools for close to two million news articles spanning a twenty year period. Researchers immediately recognized the significance of this resource. In its brief

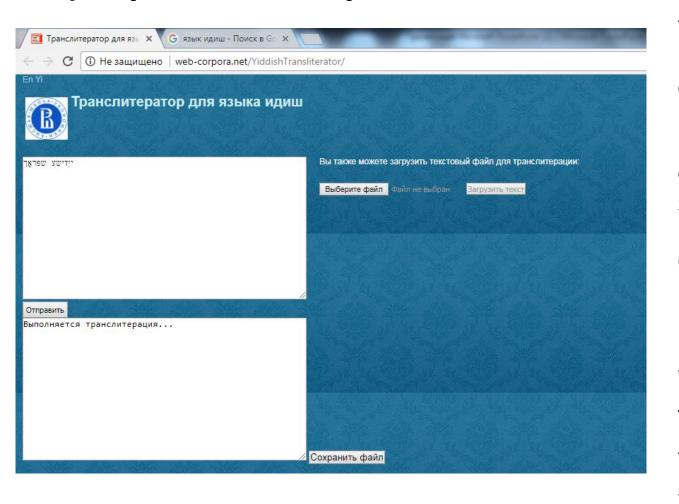
## 4) Мультимедийный корпус идиш (http://webcorpora.net/YiddishMultimediaCorpus/search/)

Язык идиш - интересный для лингвистов идиом, однако действительно лингвистических работ, посвященных этому языку, существует крайне мало, и они привлекают недостаточно фактического материала.



### МИНУСЫ: -нет инструкции

# 5) Транслитератор идиш (http://web-corpora.net/YiddishTransliterator/)



Транслитератор работает следующим образом: вы можете ввести в расположенное слева окошко любой текст на языке идиш, написанный еврейскими буквами, и нажать кнопку «отправить», после чего в поле снизу Вы увидите тот же текст в латинице, приведенный к транслитерации YIVO. Таким образом, вне зависимости от орфографии изначального текста, в транслитерации Вы сможете увидеть