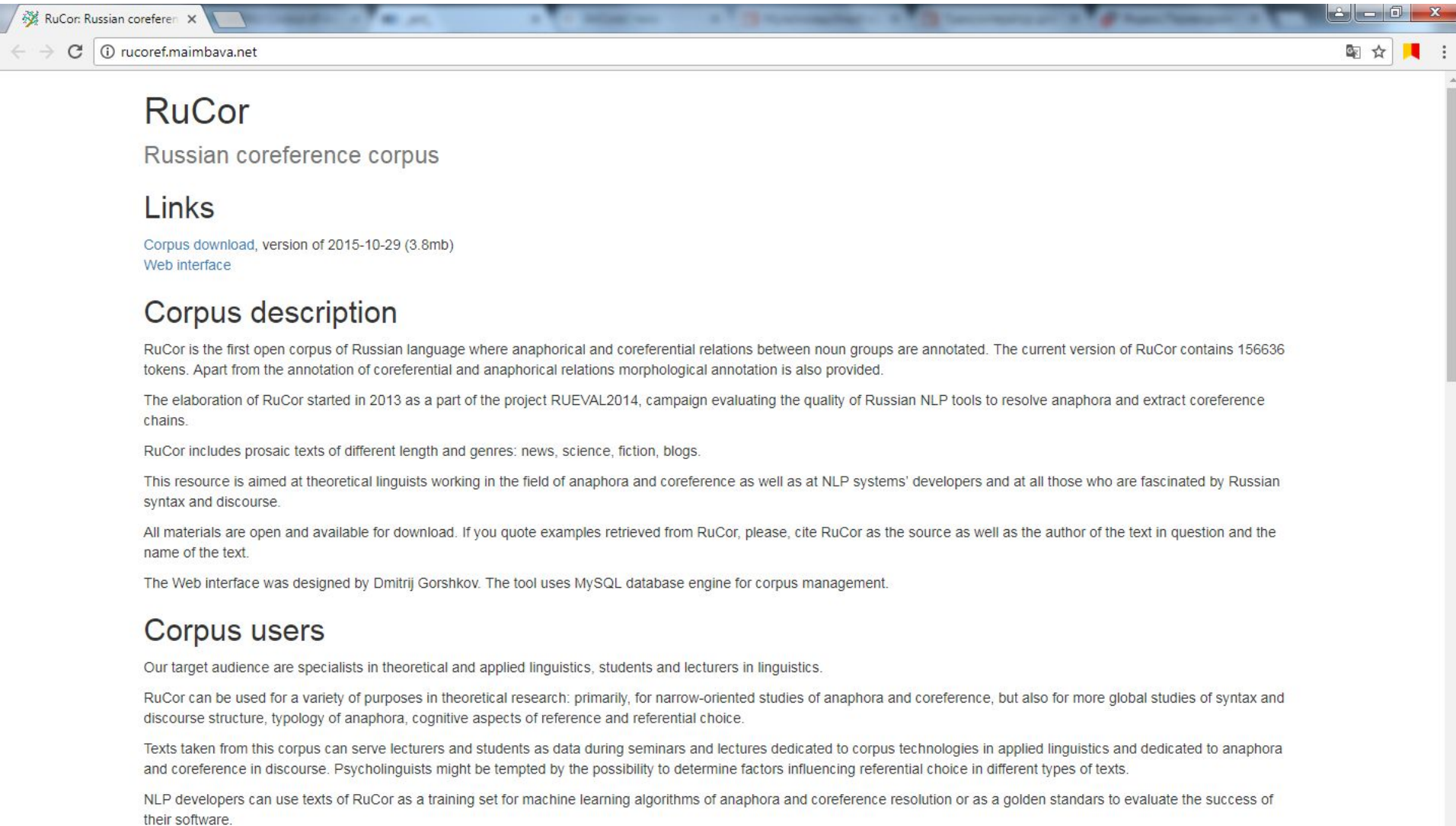


8 тема:

- 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 coreferen x

← → ↻ ⓘ rucoref.maimbava.net

RuCor

Russian coreference corpus

Links

[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	<p>1) Web формат</p> <p>2) Можно скачать</p>
Взаимосвязь в предложении	<ul style="list-style-type: none"> - анафорические и кориферентные отношения между существительными группами аннотируются - также предусмотрена морфологическая аннотация
Формат	RuCor включает в себя прозаические тексты различной длины и жанров: новости, наука, фантастика, блоги.
Аудитория	<ul style="list-style-type: none"> - 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. - специалисты в области теоретической и прикладной лингвистики, студенты и преподаватели в области лингвистики
RuCor использование	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	<p>количество текстов 181</p> <p>количество маркеров 156637</p> <p>Количество цепей корреляции 3638</p> <p>количество выбранных существительных групп 16558</p> <p>Распространение текстовых жанров:</p> <ul style="list-style-type: none"> <новости 45% эссе 21% фантастика 18% науки 9%блогов, комментариев 5% Русская Википедия 2%

Веб-интерфейс

HD_ant_ x

ant1.compling.net/res03/ant1.php?b=big

Text# 4

За кота – ответишь! Дэвид Лисс, популярный автор исторических детективов, впервые обратился к современности. И зря... [OFC/4.txt]

4. Главный герой романа Дэвида Лисса «Этичный убийца» (СПб., «Азбука-классика», 2008) – семнадцатилетний Лемюэль Алтик живет во Флориде с мамой и отчимом.

5. После окончания школы юноша получает приглашение учиться в Колумбийском университете, но отчим вопреки обещаниям отказывается оплачивать учебу пасынка.

6. Лему остается самому заработать необходимую сумму.

7. Герой выбирает сетевой маркетинг, потому что единственное, что у него получается, – это торговать подпиской на многотомную энциклопедию, убалтывая будущих клиентов.

8. Правда, когда ходишь «от двери к двери», можешь нарваться на такие неприятности, о которых и не подозреваешь...

9. Прежде чем перейти к «Этичному убийце», сделаем небольшое отступление.

10. Сорокадвухлетний американец Дэвид Лисс – специалист по английской истории, так что сюжеты первых опубликованных им остросюжетных романов отсылали читателя в Лондон первой четверти XVIII века.

11. Роман «Заговор бумага» (2000) не зря был удостоен премии «Эдгар» за лучший детективный дебют: приключенческая канва опиралась на отличное знание описываемых реалий.

12. Шпаги, деньги, биржа

13. Главный герой «Заговора бумага», бывший профессиональный боксер, а затем специалист по улаживанию всяких малоприятных дел (в наше время его бы назвали частным сыщиком), знаток жизни лондонского «дна» Бенжамен Уивер, расследуя обстоятельства смерти собственного отца, скромного еврейского маклера, обнаружил, что «несчастный случай» – тщательно спланированное убийство.

14. Причем нити заговора тянутся к финансовому монстру – «Компании южных морей», чьи махинации могут быть разоблачены изысканиями Уивера-старшего.

15. Бенжамен владел приемами кулачного боя, умело обращался со шпагой или пистолетом, а его создатель неплохо справлялся с придуманной им хитроумной детективной интригой, однако читателю в гораздо большей степени приглянулось в романе нечто другое.

Variant Select Show Edit

zc

s1

s2

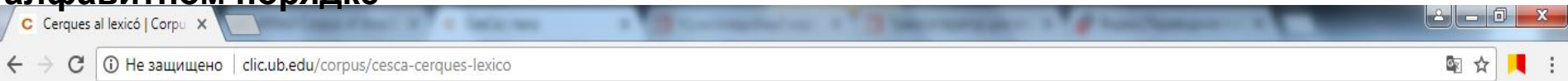
2) AnCora (<http://clic.ub.edu/corpus/>)

AnCora-это свод каталанский (анкора-CA) и испанский (анкора-ЭС) с различными

Кол-во слов	15 766 265 предложений с 537 871 550 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



Corpus

Català | Español | English

Araknion-TXT Descàrregues

Corpus

- ▶ AnCora
- IARG-AnCora
- ▼ CesCa
 - Cerques Al Corpus
 - Cerques Al Lexicó
- GRERLI
- CIInt
- ▶ NIdent
- ▶ Projecte Araknion
- ▶ Paràfrasi
- Hopinion
- SFU Review-NEG
- HistoCat
- DialCat
- Espal
- CoMIT
- ▼ Lexicons
 - AnCoraVerb_ES
 - AnCoraVerb_CA
 - AnCoraNom_ES
 - ▶ AncoraNet

Inici

Cerques Al Lexicó

Domini

Roba

Agrupar resultats

Segons anys que fa que parla català

Ordenar

Alfabèticament

Lemes del domini

- sempre (480) [(sempre) (>4) (1 a 4) (- no classificable -) (<1)]
 - abanderado (1)
 - abric (417)
 - acampanat (1)
 - accessori (4)
 - aquila de ganxo (3)
 - aquila de nit (1)
 - aleta (1)
 - all star (2)
 - amb dibuix (1)
 - americana (151)
 - amorós (1)
 - anell (96)
 - anell del peu (1)
 - anella (2)
 - anorac (302)
 - anorac de plomes (12)
 - antifac (2)
 - armadura (2)
 - armilla (279)
 - armilla antibales (1)
 - arracada (152)
 - avarca (108)
 - babutxa (4)
 - bala (1)
 - ballarina (129)
 - bandana (3)
 - bandera (1)
 - bandolera (5)
 - banvador (761)
 - barnús (120)
 - barret (562)
 - barret d'aja (1)

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:	- Английский язык - Обязательная регистрация



- ABOUT
- MEMBERS
- COMMUNICATIONS
- LANGUAGE RESOURCES
- DATA MANAGEMENT
- COLLABORATIONS

Quick Links

- CATALOG
- NEW CORPORA
- USER LOGIN
- HOW TO GET DATA
- DATA MANAGEMENT PLANS
- PROJECTS

What's New:

- LDC closed for Thanksgiving, November 23-26
- Join LDC for Membership Year 2018
- Spring 2018 Data Scholarship Program
- Web pages feature DMPs
- LDC enhances its user services
- Staff Podcasts Accessible on the LDC Blog

How LDC Data Inspires Research

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE nltf SYSTEM "http://www.nltf.org/IPTC/NITF/3.3/specification/dtd/nltf-3-3.dtd">
<nltf change.date="June 10, 2005" change.time="19:30" version="-//IPTC//DTD NITF 3.3//EN">
<head>
<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.

Словоформа: Лемма: Грамматические признаки: [выбрать](#) [задать подкорпус](#)

[Искать](#) [Очистить](#)

Записи лекций на литовском диалекте, предоставленные

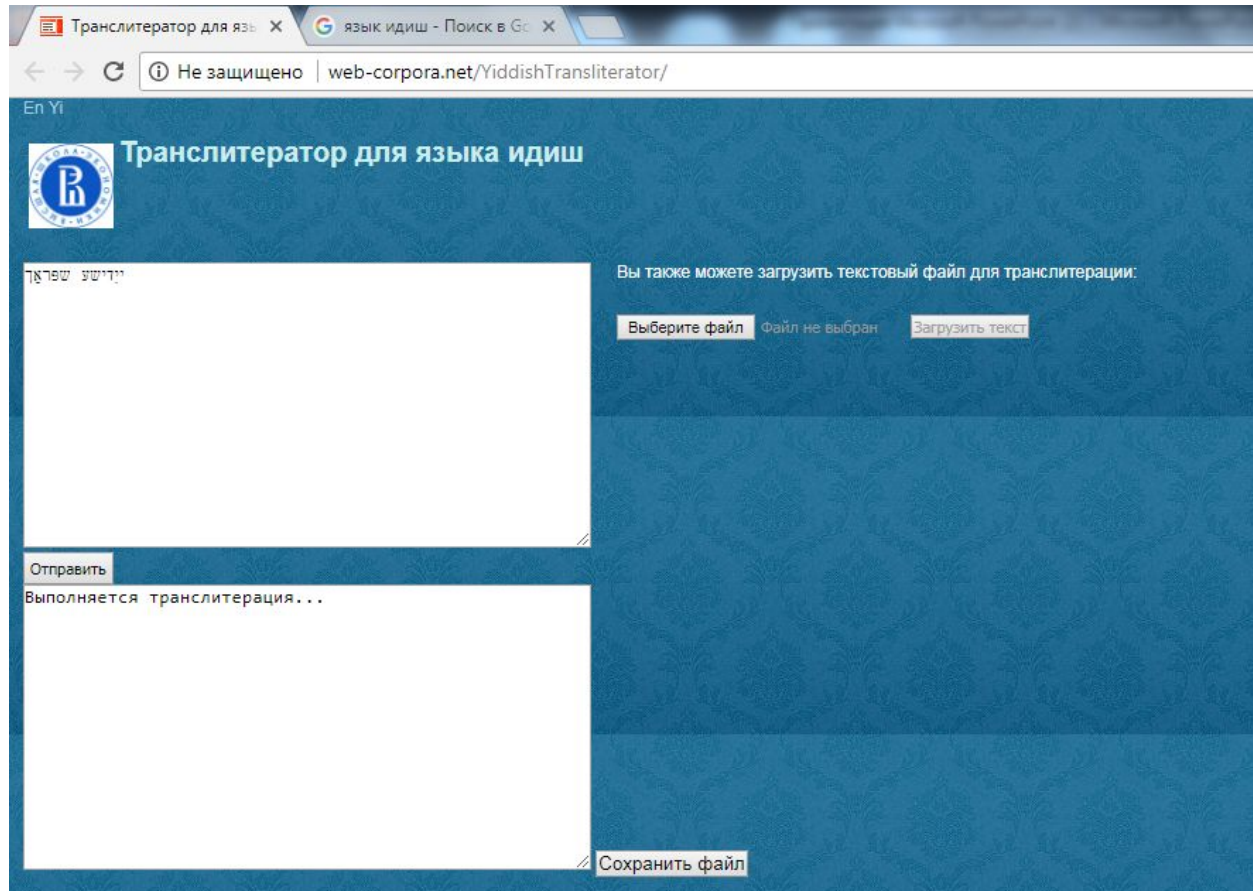
Записи фольклорных интервью на украинском диалекте из архива [ЦБИ РГГУ](#) (2007–2010 гг.)

- Интервью с женщиной из Орхея
- Интервью с женщиной из Бельц о детстве
- Интервью с парой из Рыбницы

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