8 тема:

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

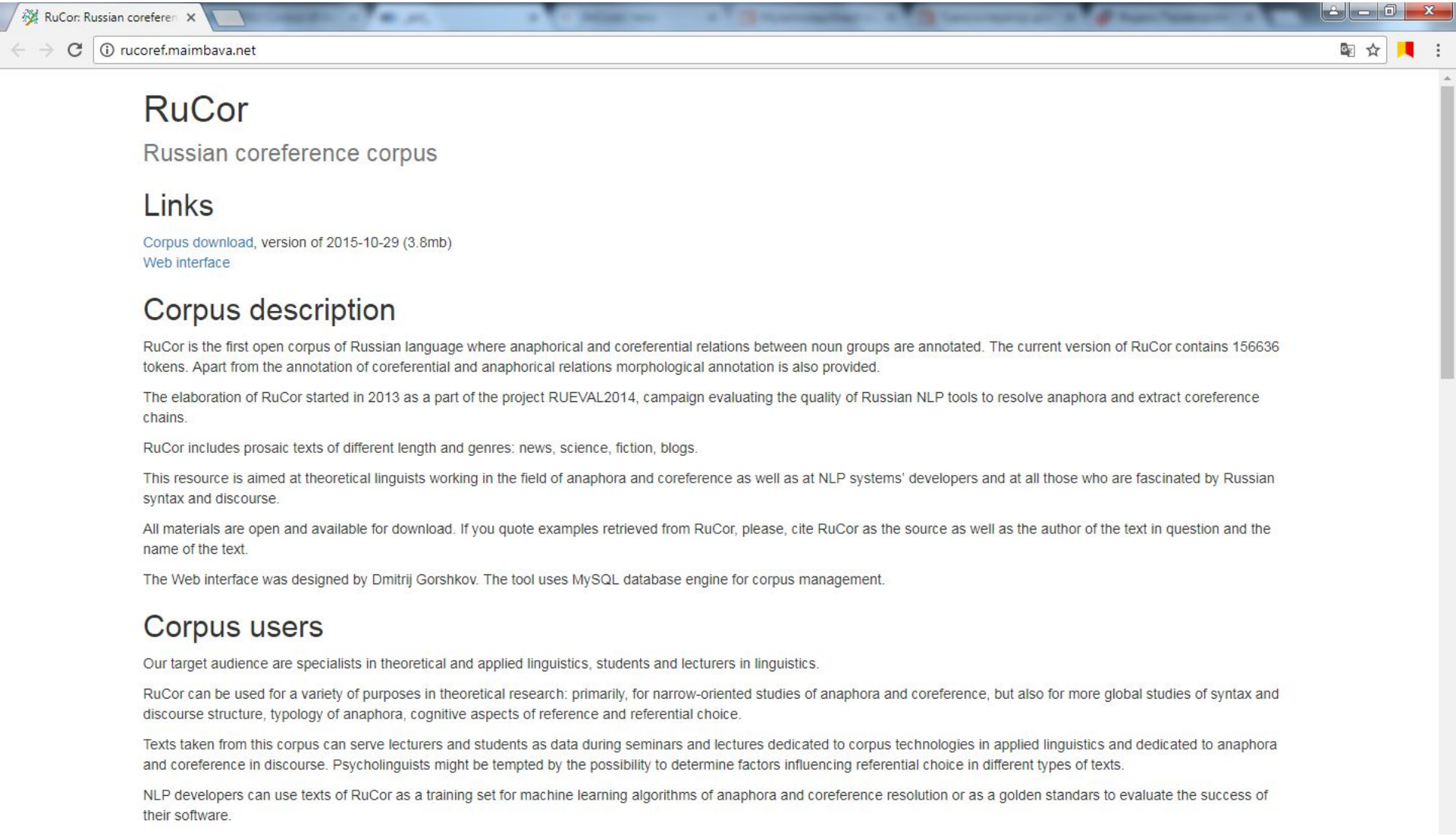
2) AnCora (http://clic.ub.edu/corpus/),

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

4) Мультимедийный корпус идиш (http://web-corpora.net/YiddishMultimediaCorpus/search/),

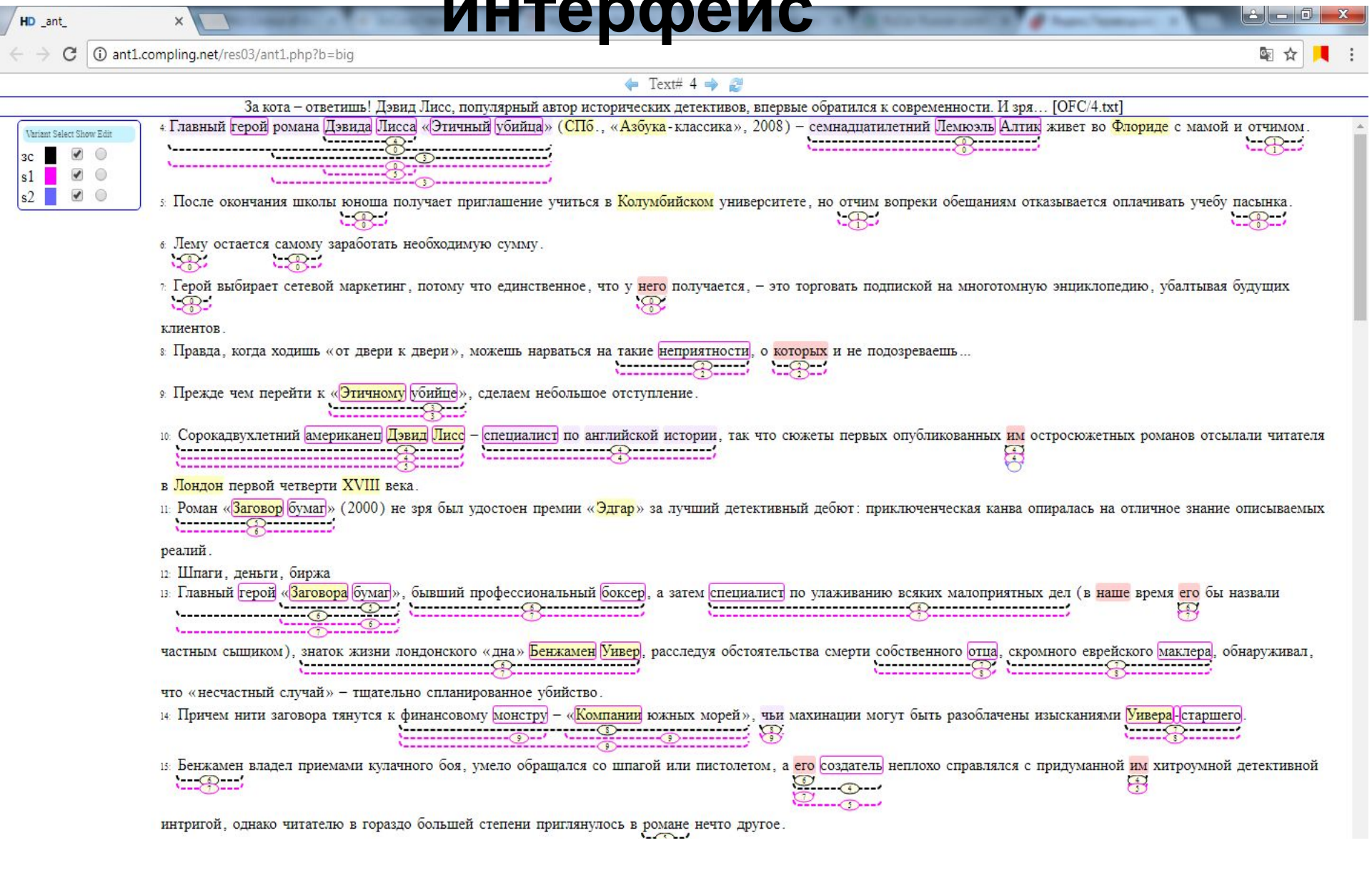
5) Транслитератор идиш (http://web-corpora.net/YiddishTransliterator/)

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



Корпус	1) Web формат 2)Можно скачать
Взаимосвязь в предложении	- 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	- 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 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



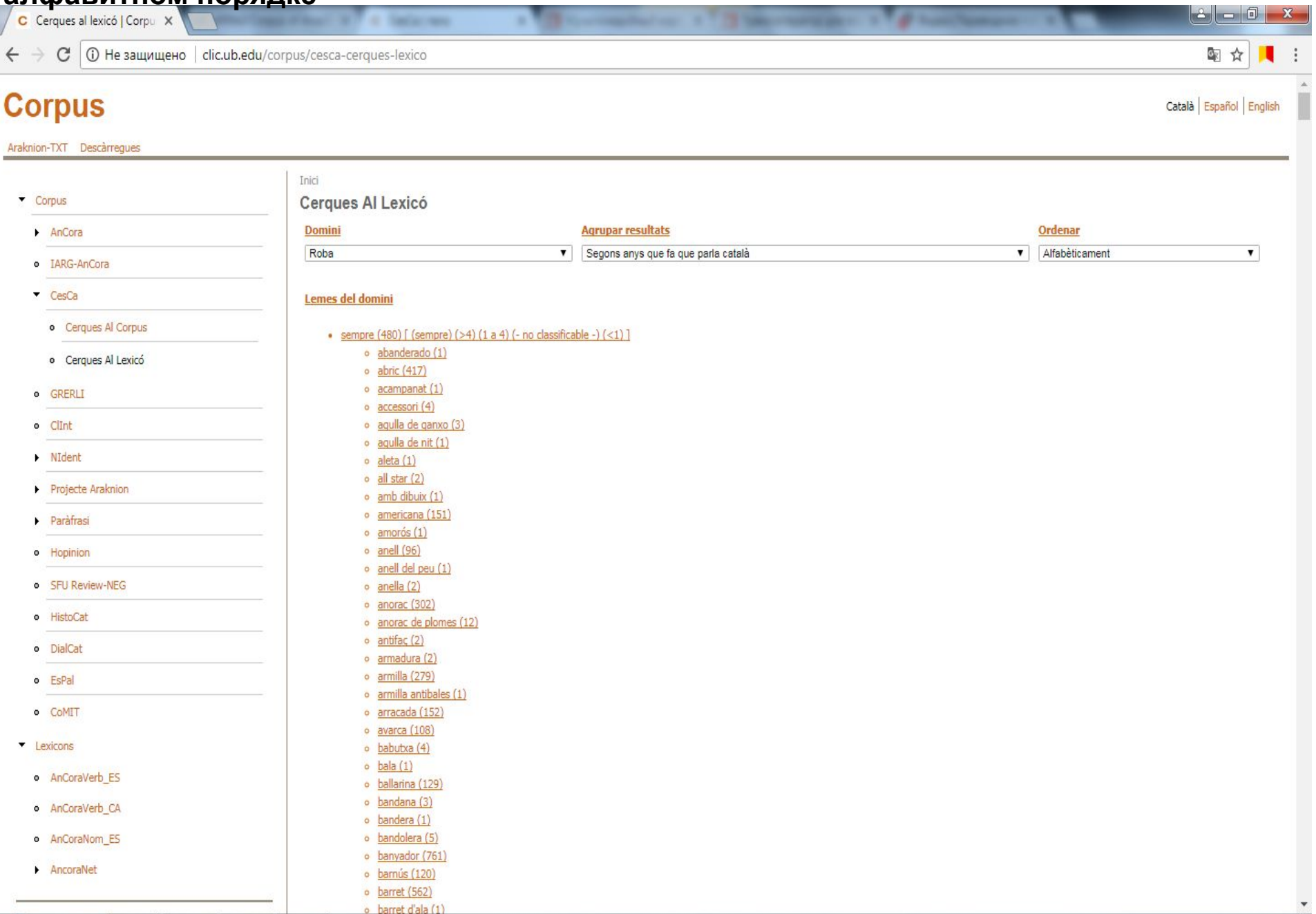
2) AnCora (http://clic.ub.edu/corpus/)

Ancora-это свод каталанский (анкора-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



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

Задания. Корпусная ли | Home | Linguistic Data C | Мультимедийный корп | Транслитератор для яз

Надежный | https://www.ldc.upenn.edu

Linguistic Data Consortium

SEARCH

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

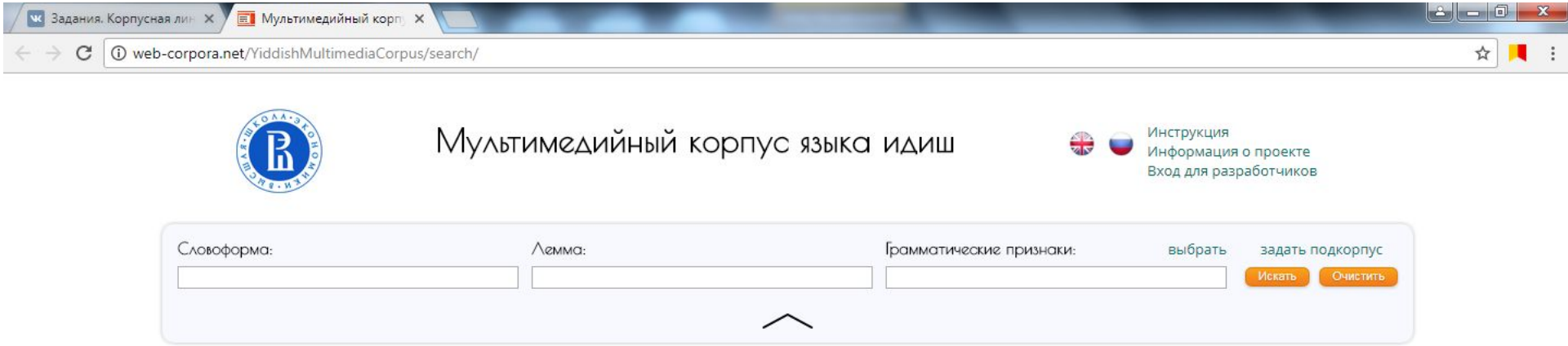
```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE nitf SYSTEM "http://www.nitf.org/IPTC/NITF/3.3/specification/dtd/nitf-3-3.dtd">
<nitf change.date="June 10, 2005" change.time="19:30" version="-//IPTC//DTD NITF 3.3//EN">
 <head>
  <title>Irabu, at His Best, Silences Favorite Foe</title>
  <meta content="02YANK$03" 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/)

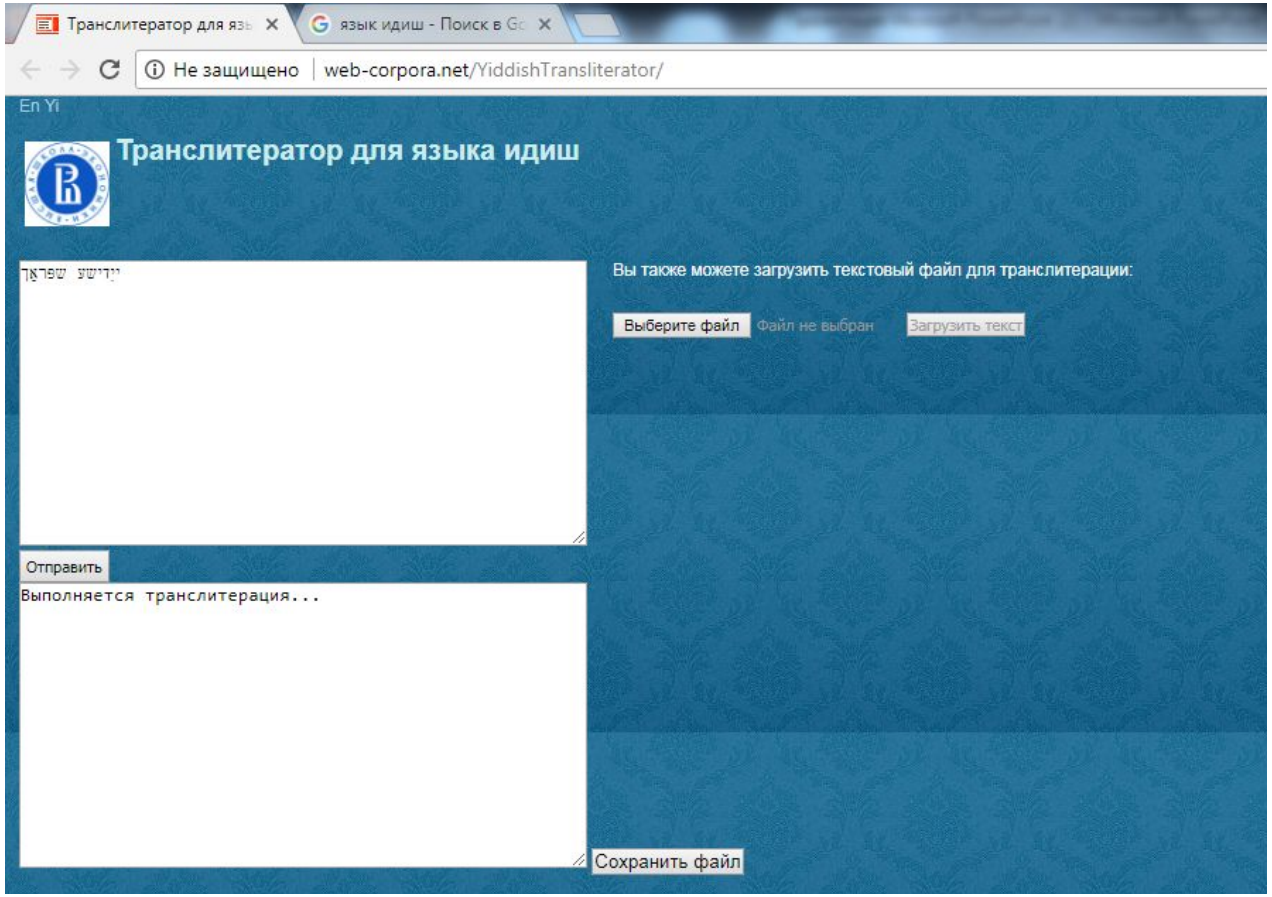
Язык идиш - интересный для лингвистов идиом, однако действительно лингвистических работ, посвященных этому языку, существует крайне мало, и они привлекают недостаточно фактического материала.



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