Корпусные технологии.

В информационную эру развития компьютерных технологий и широкой доступности интернета открываются новые возможности для повышения эффективности процесса обучения иностранным языкам. Одним из современных и актуальных направлений в методике преподавания иностранных языков является обучение разным аспектам языка на базе корпусов. Корпусные технологии используются при обучении лексике, грамматике, переводу и т.д.

Рассмотрим такие корпусные технологии, как конкордансеры и корпусные менеджеры.

Конкорданс – это список всех употреблений заданного языкового выражения в контексте со ссылками на источник (в этом значении используется в корпусной лингвистике).

Поиск в корпусе данных позволяет по любому слову построить конкорданс. Обычно **конкордансером** называют список примеров, полученных в результате поиска по корпусу. Иными словами, **конкорданс** – список контекстов, где искомая единица представлена в ее лексическом окружении и характеризуется набором статистических данных.

Конкордансы представляют собой специальные программы, предназначенные для обработки текста с той или иной лингвистической задачей, заключающейся в поиске морфем, слов и словосочетаний в контексте. Например, в конкордансе можно отследить в группе текстов варианты использования какой-либо грамматической конструкции. И программа выдаст все примеры заданной грамматической конструкции вместе с контекстом. Также, благодаря репрезентативности, примеры позволяют проводить собственные исследования, прослеживая изучаемые лексические и грамматические явления в языковых контекстах.

Лингвистические **задачи**, для решения которых используют конкордансы:

- сравнение различных использований одного слова

- анализ ключевых слов

- анализ частотности слов и словосочетаний

- поиск и исследование фраз и идиом

- поиск перевода

- создание списков слов

Конкордансеры позволяют получать частоту той или иной языковой единицы по произвольному корпусу текстов, список контекстов, в которых данная единица встретилась. Многие из них позволяют сортировать контексты по ключевому слову или по словоформе, по ближайшему контексту.

Программы-конкордансеры: Concordance; Micro-Concord; MonoCorc; TACT (Text Analysis Computing Tools); TACTWeb; SARA и др.

Корпусный менеджер – специализированная поисковая система управления текстовыми и лингвистическими данными, которая включает программные средства для поиска в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме.

Характеристики корпусного менеджера (В.П. Захаров):

- формальная релевантность

- информационно-поисковый язык фактографического типа

- **НО**: умение работать с лексемами и словоформами

- операции над запросами

- сравнение с архитектурой поисковых систем в сети Интернет: роботы, программы загрузки индексов, поисковые системы.

Таким образом, корпусный менеджер стоит на порядок выше конкордансера. Если к задачам конкордансера можно отнести построение конкорданса отдельных слов, словосочетаний, частей слов, знаков пунктуации и т.д., то корпусные менеджеры способны строить полные конкордансы, включающие в себя не только слова, но и другие элементы корпуса.

Корпусные менеджеры предназначены для работы с такими явлениями, как лемма и морфологические характеристики слова, позиция слова в предложении ив структуре размеченного текста, библиографические и типологические признаки документа, статистические данные и др.

Программы корпусных менеджеров: Bonito; CQP; DDC; WebCorp; Xaira.

Еще несколько важных **понятий**:

Частотность – процент искомой единицы от числа соответствующих единиц (слово относительно всех слов, биграммы относительно всех биграмм и т.д.).

Относительная частота употребления (ipm) для данного слова определяется количеством употреблений этого слова за год, поделенное на объем корпуса за этот год и умноженное на 1 миллион.

Корпусные приложения.

NGram Viewer

N-грамма — последовательность из n слов.

Сервис **NGram** перебирает базу оцифрованных книг Google и позволяет выявить частоту употребления слов и фраз в книгах в разные исторические периоды.

Лингвистическую программу на платформе поисковика Google запустили гарвардские ученые. Они закачали на сервер около 5 миллионов книг, опубликованных за период с 1800 по 2008 год. Любой желающий теперь может проследить, с какой частотой определенные слова упоминались на протяжении веков.
Поиск доступен на семи языках: английском, французском, немецком, испанском, иврите, русском и упрощенном китайском. Особое место отведено английскому — он здесь подразделяется на «общий», «художественны», «британский» и «американский». Система позволяет проводить сопоставительный анализ — к примеру, Ленин упоминается в литературе на русском языке гораздо чаще, чем Сталин, в том числе и после 2000 года, что легко можно проследить на графике, если ввести эти фамилии через запятую в строку поиска. По этой причине сервис привлек внимание серьезных исследователей — человеку и целой жизни не хватит, чтобы проанализировать 5 миллионов источников, а компьютерная программа проделывает эту операцию за несколько секунд. Под онлайн-сервис в Гарварде придумали целую науку, которую решили назвать «культуромикой». Дело в том, что отсканированные книги в основном посвящены вопросам культуры и общества, здесь нет технической литературы.

AntConc

С помощью данной программы можно производить поиск и подсчет различных элементов текста, анализировать частотность и контекст употребления словоформ, словосочетаний и морфем, сравнивать употребительность словоформ в разных текстах.

Отсутствие морфологического анализатора частично компенсируется возможностью подключения пользовательского списка лемм. Программа

может быть использована для получения привязанных к заданной предметной области словарных минимумов, списков устойчивых сочетаний (в том числе терминологических), выборок к тематическим группам слов.

Проще говоря, это программа, которая позволяет создать собственный корпус.

Sketch Engine

Система **«Sketch Engine»** является веб-системой, которая позволяет лингвистам исследовать большие корпуса текстов и создавать сложные запросы, чтобы извлекать нетривиальную информацию из этих корпусов.

Система содержит 292 готовых текстовых корпусов, которые пользователь может использовать для своих исследований. Если рассматривать количество корпусов по языкам, то всего используется 70 языков. Если посчитать количество слов во всех корпусах, то получим 240 279 265 530 слов. Наибольшим из них является корпус английского языка «enTenTen», который состоит из 19 717 205 676 слов и 22 878 431 750 токенов, самым маленьким является корпус африканского языка «CHILDES Afrikaans Corpus», который содержит 26 020 слов и 33 134 токенов.

Данная система позволяет загрузить собственный корпус текстов в различных форматах: «doc», «docx», «htm», «html», «pdf», «ps», «tar.bz2», «tar.gz», «tgz», «tmx», «txt», «vert», «xml», «zip». После того, как файлы были загружены, корпус необходимо откомпилировать: создать схемы слов, словарь для слов, выделить термины, выделить память для хранения данных, очистить предыдущее хранилище. К счастью, данные действия система производит автоматически, а пользователь видит только процентную шкалу, на которой отображается каждый из процессов.

Для анализа текстов пользователю доступен большой спектр инструментов, начиная от обычного поиска слова в тексте, заканчиваю специальными фильтрами для поиска предложений по определенной схеме, также система имеет свой собственный регулярный язык, который позволяет пользователю, находи определенные типы предложений и создавать различные специализированные запросы. В системе не существует возможности для сохранения полученных результатов на компьютер пользователя. Система разграничивает функциональные возможности по ролям, некоторая часть корпусов текстов является закрытой для использования. Также имеется ограничения на создание собственных корпусов текстов. Система не имеет программного доступа и не может сохранять динамически добавляемые атрибуты.

система является хорошей для проведения исследований на больших текстовых корпусах. Система обладает большими функциональными

возможностями для анализа текстовых корпусов, имеет динамически расширяемое хранилище и готовые текстовые корпуса. Однако данная система не является полностью бесплатной и предоставляет исследователю бесплатно только хранилище на 1.000.000 слов и 30 дней бесплатного использования.

Word2vec

Word2vec — программный инструмент анализа семантики естественных языков, представляющий собой технологию, которая основана на дистрибутивной семантике и векторном представлении слов. Этот инструмент был разработан группой исследователей Google в 2013 году. Работу над проектом возглавил Томаш Миколов.

Параллельные корпуса.

Параллельный корпус (Parallel Corpora) - это электронный аналог параллельных переводных текстов, как правило состоящий из множества блоков "текст-оригинал и один/несколько его переводов". Электронные тексты в корпусе могут представлять собой целое оригинальное словесное произведение или какую-либо его часть.

В современной корпусной лингвистике выделяется два вида параллельных корпусов:

1) многоязычный, или Comparable (Multilingual) Corpora,

2) переводной, или Translation Corpora.

Структурная организация корпуса может быть самая разная, в зависимости от прагматических целей его создателя или пользователя:
- в виде традиционного текста со ссылкой на переводы,
- в табличной "зеркальной" форме, что более удобно для восприятия и сравнения,
- в виде базы данных.

Направления корпусной лингвистики, в том числе проекты электронных корпусов текстов, активно развиваются и имеют значительный прикладной потенциал в методике обучения иностранным языкам и переводу, а также в компьютерной лингвистике.

В задачах обучения переводу параллельные корпусы текстов могут рассматриваться как реферативная информация и предоставлять образцы профессионального перевода при изучении приемов и способов перевода. В задачах обучения иностранному языку такие корпусы позволяют подобрать

возможные эквиваленты изучаемой лексики, проследить ее значения и функции в тех или иных контекстах.

В настоящее время особенно распространены корпусы (или параллельные тексты) художественной литературы, хотя для обучения переводу в вузе следует разрабатывать корпусы разных жанров и стилей и в первую очередь ориентироваться на научно-технические, публицистические и деловые тексты.

Анализ корпусов текстов, методы и наработки корпусной лингвистики являются перспективным направлением в области преподавания иностранных языков. Мировая практика развития этой области доказывает эффективность такого рода приложений, хотя в настоящее время возможности методов корпусной лингвистики в России пока не находят должной реализации в прикладной лингвистике, лингвистическом обучении, обучении родному и иностранному языку.

Поэтический корпус.

Поэтический (под)корпус – часть НКРЯ со специфической *метаразметкой*, в которой отражены основные жанровые и формальные параметры поэтического текста.
Присутствие данной метаразметки позволяет программными средствами восстановить акцентную схему каждой входящей в рассматриваемый поэтический текст словоформы с определенной точностью. Несмотря на то, что диахрония и синхрония русского ударения более чем подробно рассмотрены в работе, использования поэтических текстов позволяет существенно уточнить некоторые отдельные факты. В историческом аспекте это особенно важно из-за скудости акцентуированного языкового материала XVIII-XIX вв.; в аспекте современного состояния русской акцентной системы это позволяет зафиксировать факты, характерные уже не только для современного разговорного языка, но являющиеся частью формирующейся литературной нормы. Также становится возможным проследить судьбу некоторых относительно распространенных акцентных архаизмов или варваризмов.
Рассмотрим представление языковой информации в поэтическом корпусе. Интерфейс поиска, в целом, одинаков для поэтического и основного корпусов НКРЯ. Однако в поэтическом корпусе существует дополнительный набор метатекстовых атрибутов, позволяющих осуществлять поиск по характерным параметрам поэтического текста. Кратко рассмотрим те параметры, которые принципиальны для нашей работы:
· *Метр*. Традиционные силлабо-тонические метры: ямб, хорей, дактиль, амфибрахий, анапест.

Силлабические метры обозначаются, напр., *С12ж* в строке запроса или указанием параметра «силлабический» в меню.

Тонические метры представлены строчными логаэдами различного вида, дольником, тактовиком и акцентным стихом; также возможно задание обобщенной характеристики метра — «тонический».

Кроме того параметр «метр» может принимать значения «*свободный стих*» и «*гетерометрия*». Специально отмечаются гекзаметр и пентаметр как наиболее традиционные для русской поэзии типы дольника (ср. [Гаспаров 2001 : 141], где проясняется почему гекзаметр и его дериваты принято относить к тоническим метрам[2]).

· *Стопность*. В меню представлены наиболее распространенные типы стопности: однородная (2, 3, 4-ст. и т.д.), вольная (3,4; 4,6-ст. и т.д.), урегулированная (4+2, 4+3-ст. и т.д.). В строке запроса допускается ввод значений типа 4(5), по которым отбираются 4-ст. стихотворения со спорадически возникающими 5-ст. строками. Этот параметр действует и для тонических (означает количество иктов) и для силлабических стихов (означает количество слогов); не применяется к верлибрам.

· *Клаузула*. В меню указаны наиболее частотные типы клаузул, однако (особенно это актуально для сложной строфики) пользователь также может указать нужное чередование в строке запроса. Также возможен поиск по обобщенному типу клаузулы — вольной или регулярной — без дополнительных уточнений.

Более подробно особенности этих параметров в контексте нашей темы будут рассмотрены ниже.

Силлабо-тонические и тонические тексты поэтического корпуса снабжены обозначениями сильных мест стопы (иктов) в соответствии с типом метра. Для силлабо-тонических текстов это (за редкими исключениями) позволяет указать одну из возможных позиций ударения в словоформе. Для тонических текстов, однако, надо иметь ввиду возможную вариативность ударения, ограниченную, однако, требованиями соблюдения метрической схемы (особенно строгой в дольнике). Для силлабических текстов ударение фиксируется только в зоне клаузулы, т.к. убедительно судить об акцентуации форм из этих текстов нельзя.

Метрическая разметка дополняется морфологической разметкой, данные которой позволяют сравнить реально наблюдающуюся акцентуацию словоформы (в метрической схеме) и теоретическую, основанную на данных русской грамматики. Наиболее любопытны, конечно, случаи расхождения этих двух параллельных источников, т.е. случаи, когда, пользуясь техническими терминами стиховедения, ударение падает не на сильное, а на слабое место в строке.

Лабораторная работа.

Google Books NGram Viewer

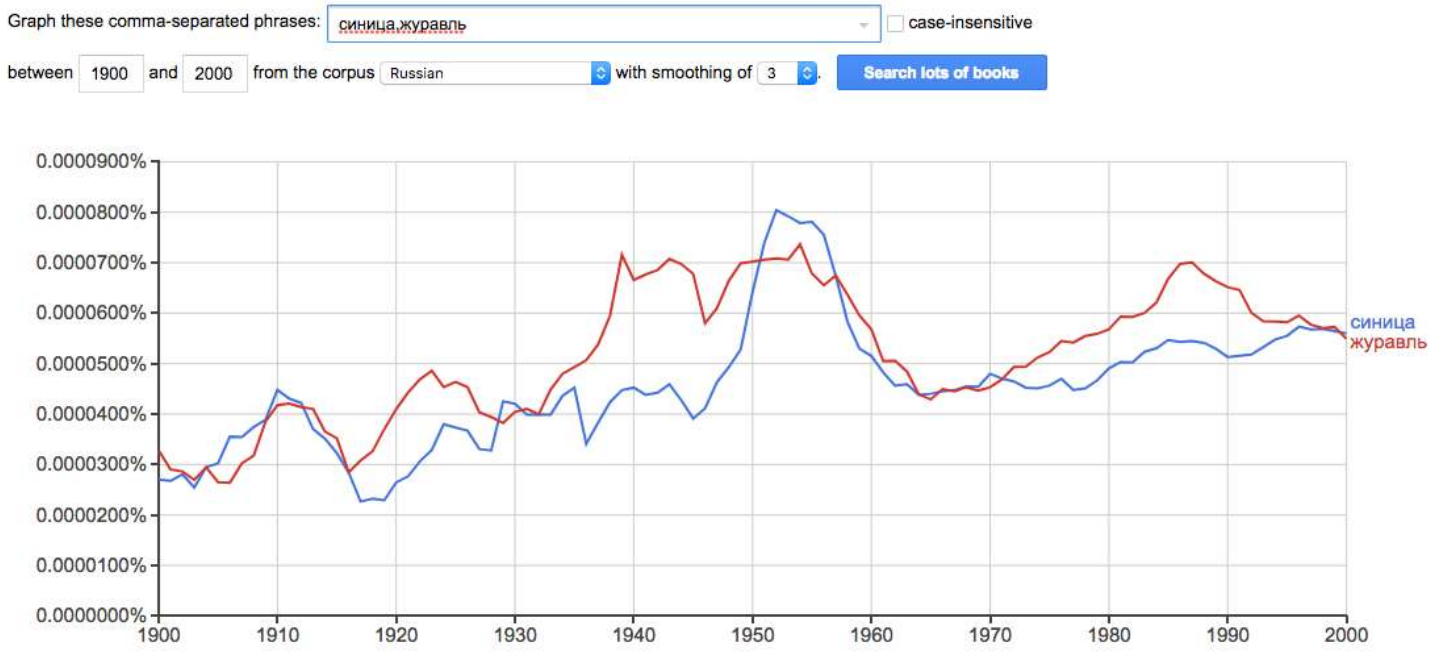
Обозначения:

NOUN	
VERB	
ADJ	adjective
ADV	adverb
PRON	pronoun
DET	determiner or article
ADP	an adposition: either a preposition or a postposition
NUM	numeral
CONJ	conjunction
PRT	particle
ROOT	root of the parse tree
START	start of a sentence
END	end of a sentence

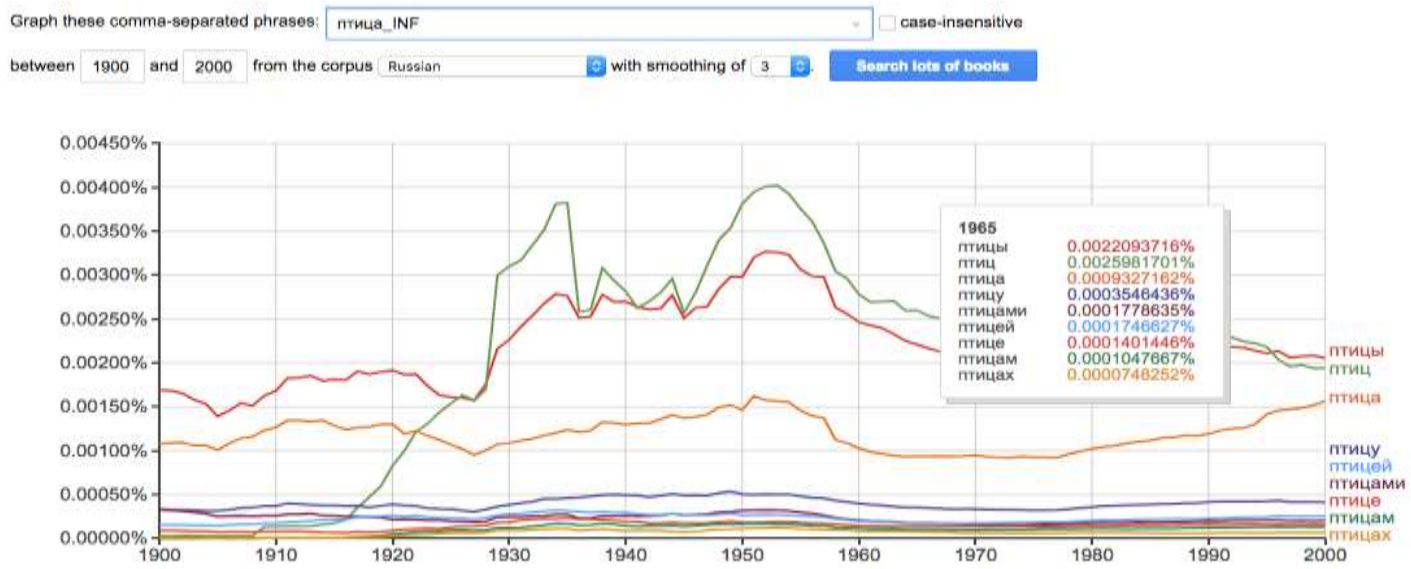
Informal corpus name	Shorthand
American English 2012	eng_us_2012
American English 2009	eng_us_2009
British English 2012	eng_gb_2012
British English 2009	eng_gb_2009
Chinese 2012	chi_sim_2012
Chinese 2009	chi_sim_2009
English 2012	eng_2012
English 2009	eng_2009
English Fiction 2012	eng_fiction_2012
English Fiction 2009	eng_fiction_2009

https://books.google.com/ngrams/graph?content – ссылка на сайт.

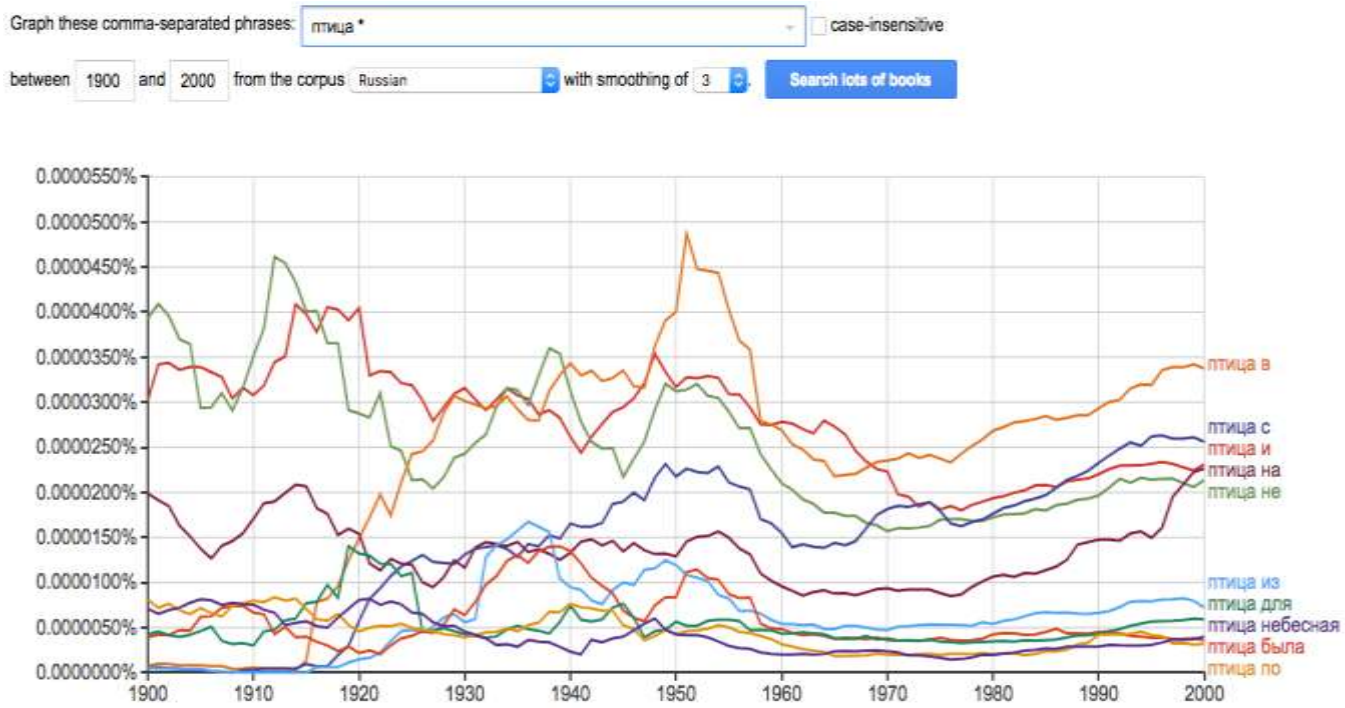
Чтобы получить сравнить частотности нескольких единиц, запишите их через запятую.



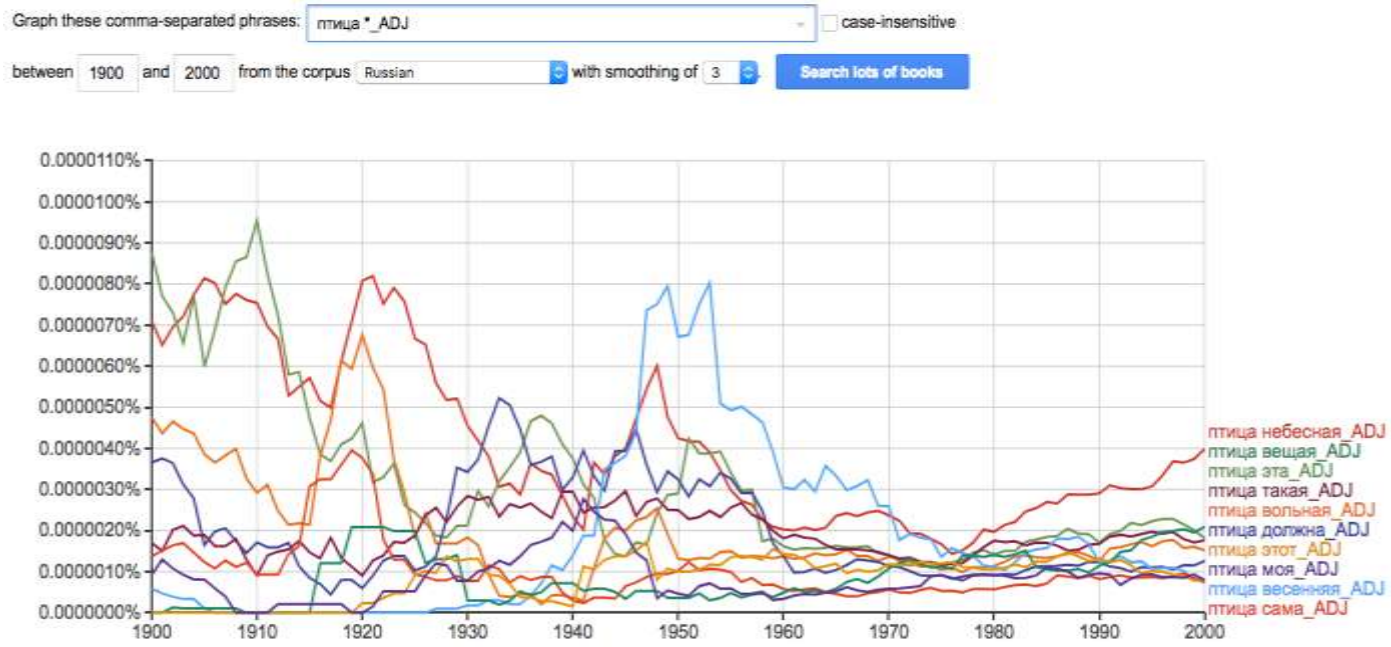
По умолчанию осуществляется поиск конкретных словоформ (как *Точный поиск* в НКРЯ), если вы хотите искать все словоформы, припишите _INF в конце слова (например: птица_INF).



Если вместо одного из слов поставить астериск, то буду показаны 10 самых частотных биграмм со вторым словом.



Искать также можно и грамматические характеристики.



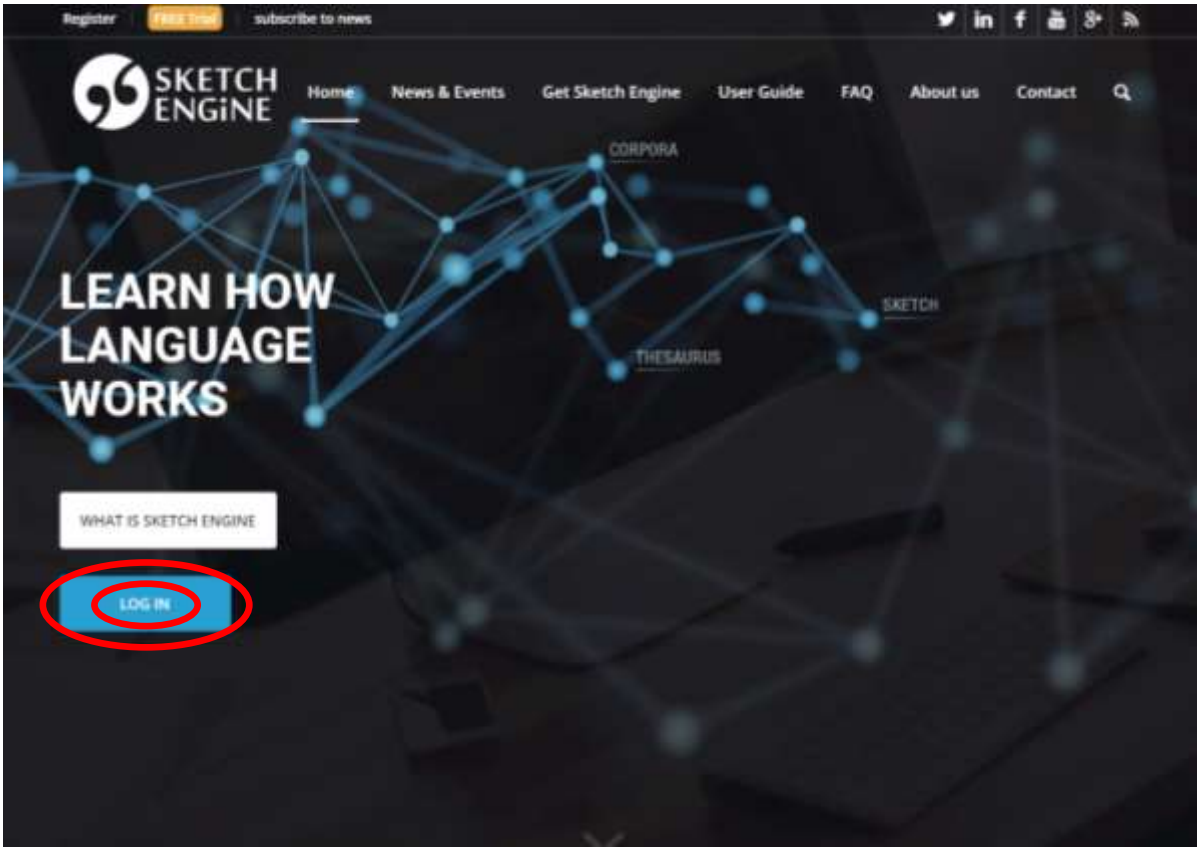
Задания:

1. Сравните частоту употребления слов «Сталин» и «Ленин».
2. Слово «grade» характерно для американского английского, а слово «form» характерно для британского английского. Выясните, так ли это?

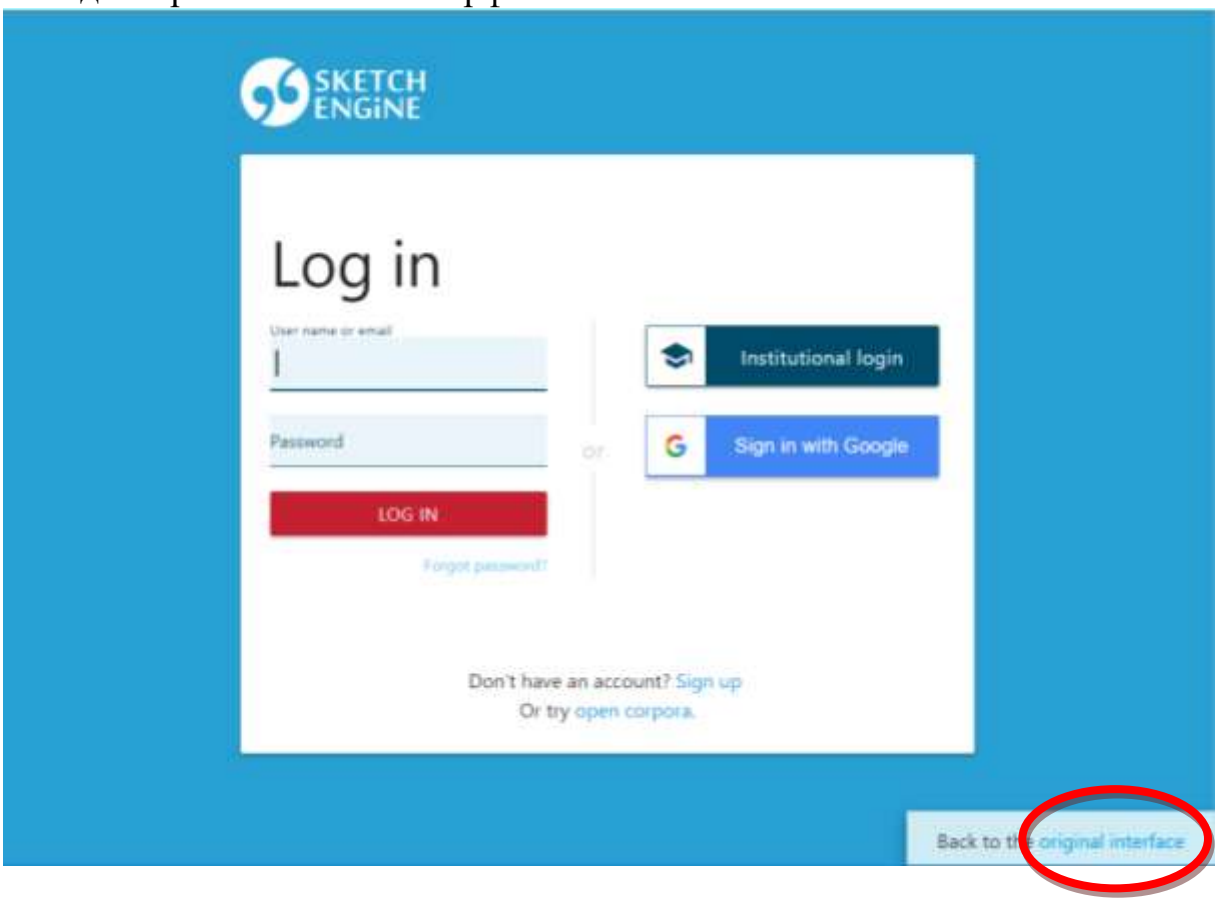
SketchEngine

Ищем в поисковике «SketchEngine».

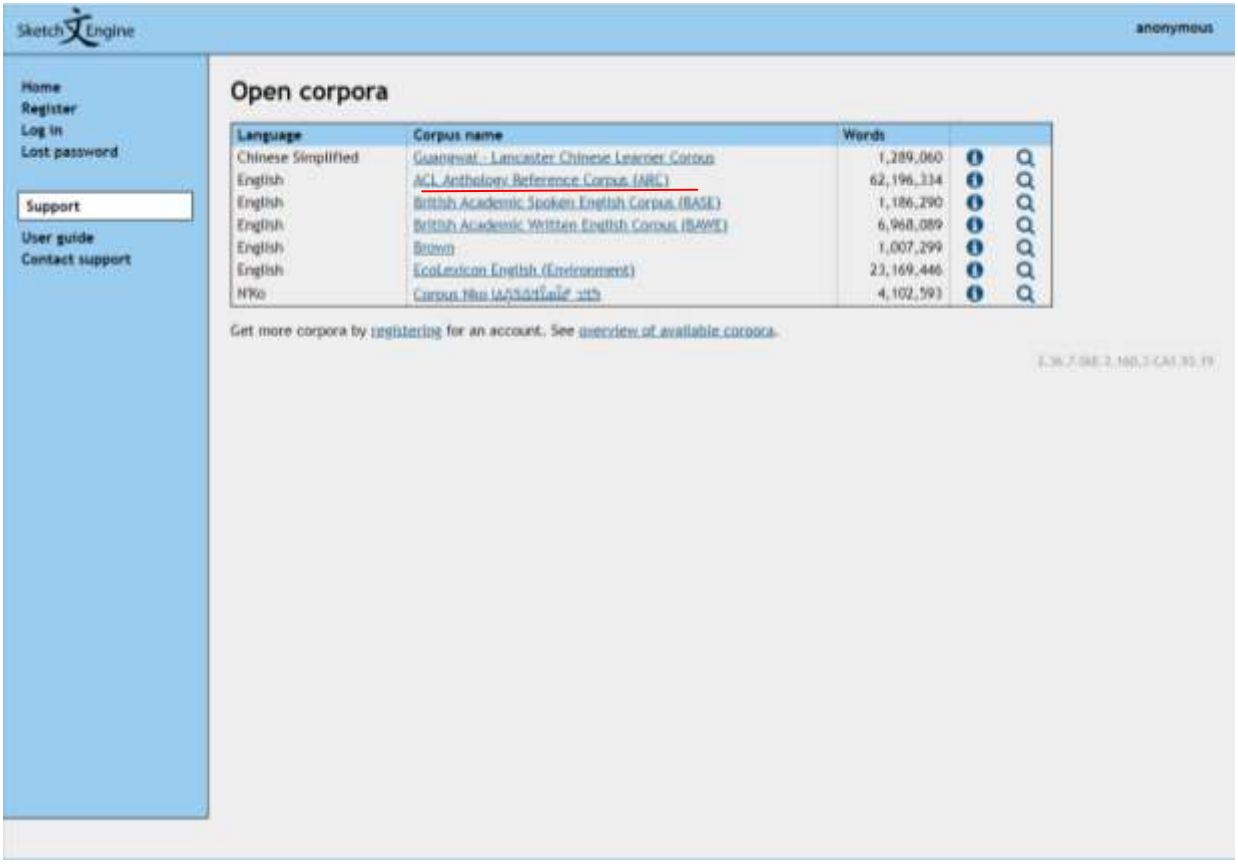
Мы попадаем на главную страницу. Нажимаем на «Log in».



Находим оригинальный интерфейс.



Нажимаем кнопку «Home», в таблице открываем вторую с верху ссылку.



1 задание. Во вкладке «Thesaurus» вводим слово «form».
2 задание. Во вкладке «Sketch diff» ищем разницу между словом «form» и словом «grade».

AntConc

Download AntConc

Открываем сайт НКРЯ. В поиске ищем слово «язык». Копируем первые три страницы в Word и сохраняем в одном из следующих форматов: .txt/.xml/.html. Дальше скачиваем **AntConc** и выполняем следующие действия:

1.Открываем во второй сверху строке меню кнопку «Word List» (вторяя слева) и нажимаем кнопку «Start» (внизу ближе к левому краю). Программа выстроит все словоформы текста в порядке частотности

2.Можно сортировать и по другим критериям. Если вместо «Sort by Freq» (в самом низу) выбрать «Sort by Word», произойдет сортировка по алфавиту, если выбрать «Sort by Word End», сортировка пойдет по концу слов.

3.Если к тому же поставим галочку между фразами «Sort by» и «Invert Order», то сортировка пойдет в обратном порядке — от редких слов к частым или от *я* до *а*.

4.Можно кликнуть из списка любое слово, начнется его автоматический поиск в окне *Concordance*.

https://quizlet.com/354791080/flash-cards/ - здесь вы можете проверить, как вы запомнили важные термины из этого урока.
https://docs.google.com/forms/d/e/1FAIpQLSdCMmS8gE4hY.. - здесь можно пройти тест
Самостоятельно просмотреть ресурс «**Word2vec**» и сделать скриншот (RusVectōrēs).
AntConc 3.2.4 Tutorial 1: Concordance Tool - Basic Features – здесь можно посмотреть мастер-класс по AntConc от Лоренца Энтони.