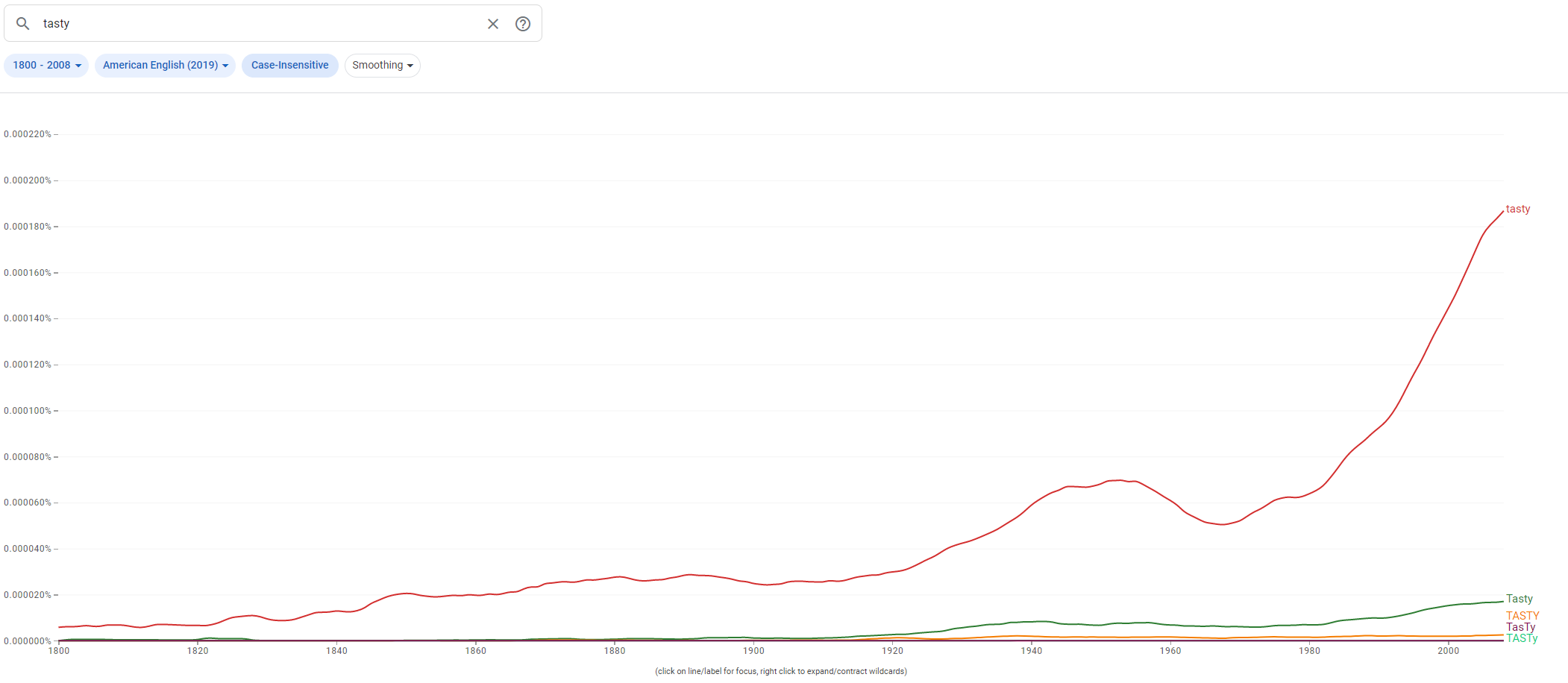
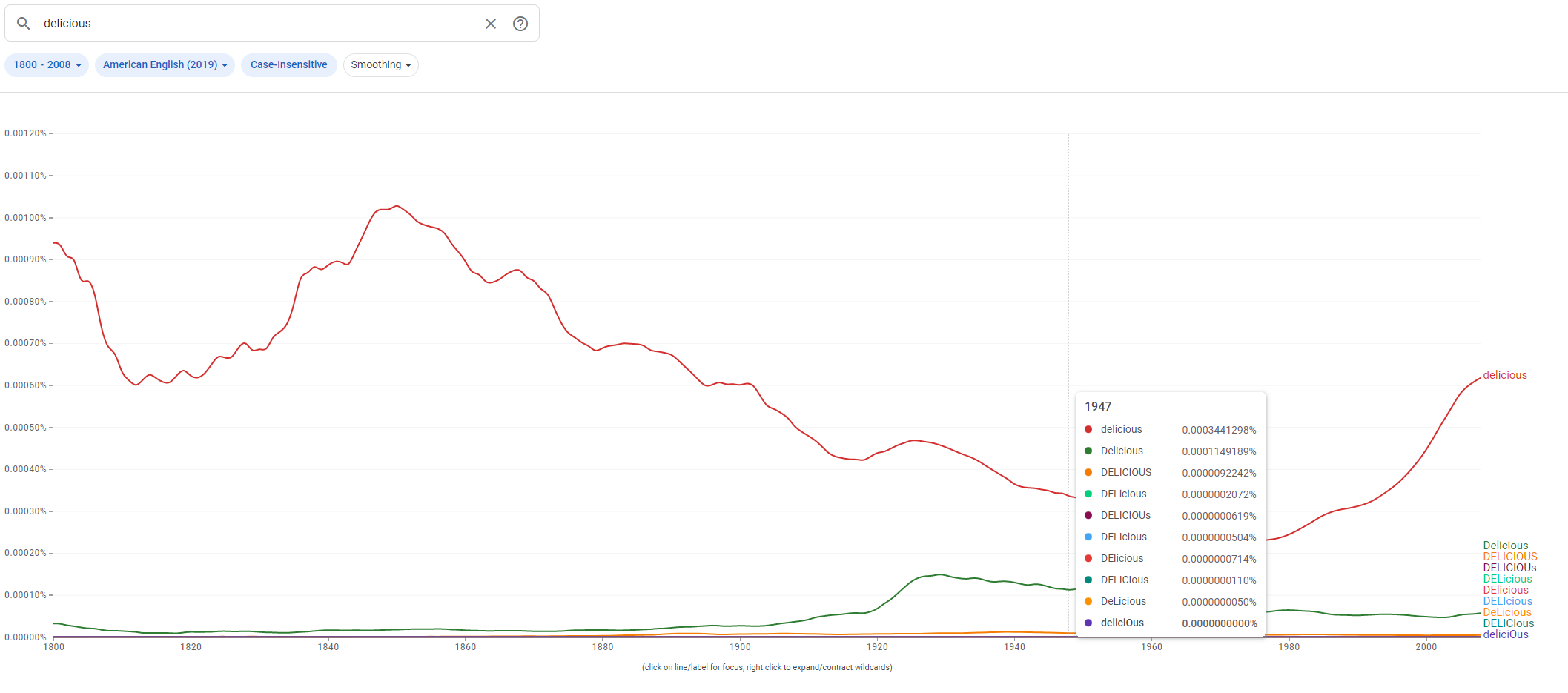
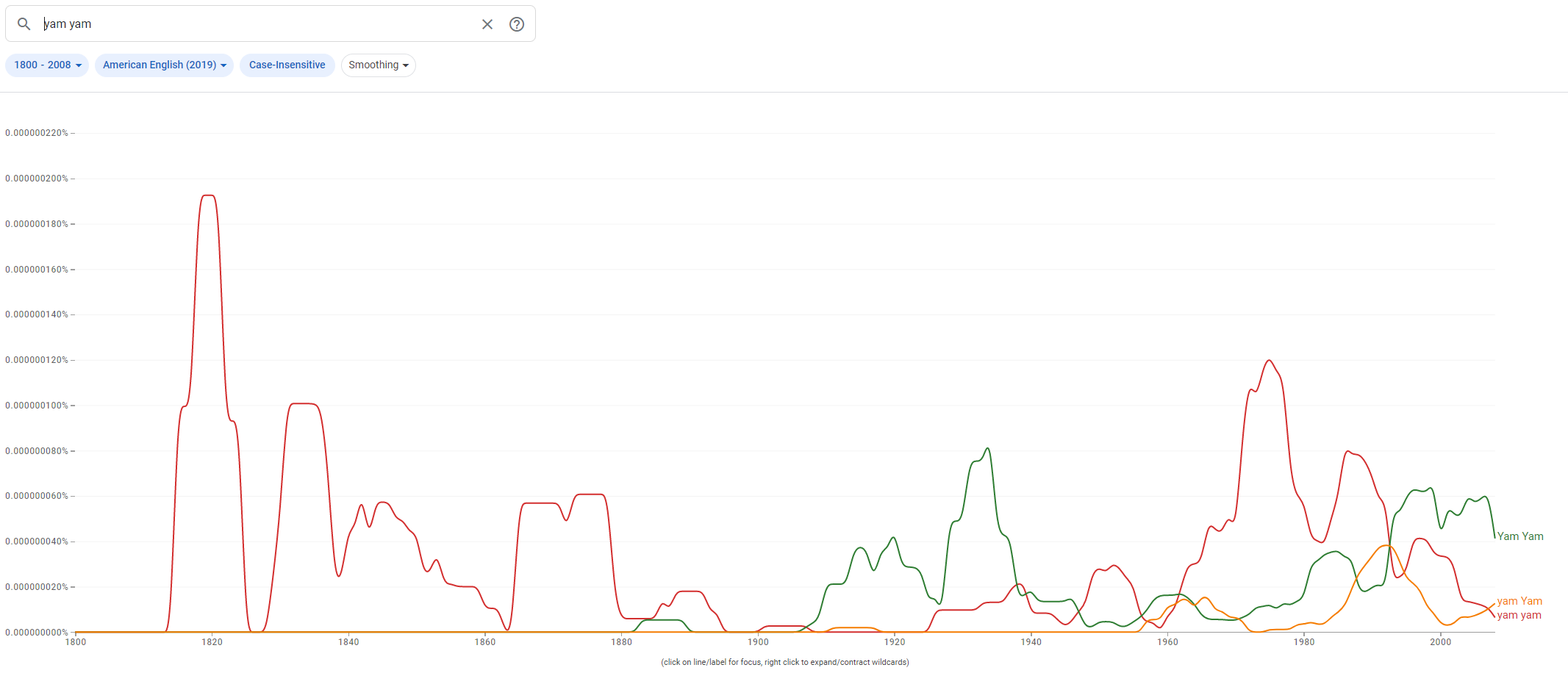
**Уровень 1**

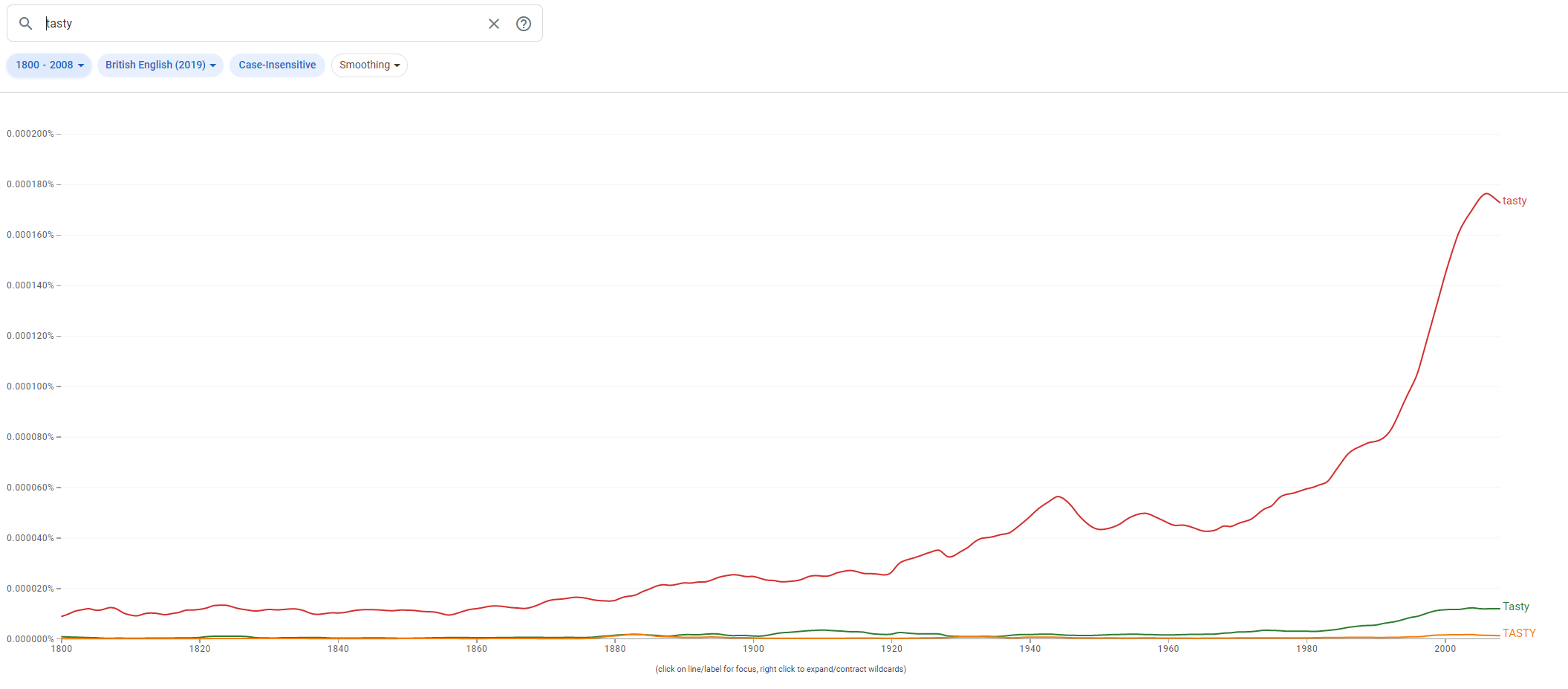
Примеры запросов tasty, delicious и yam yam соответственнов корпусе американского английского с 1800 по 2008 г. с параметром smoothing (сглаживание) 3:

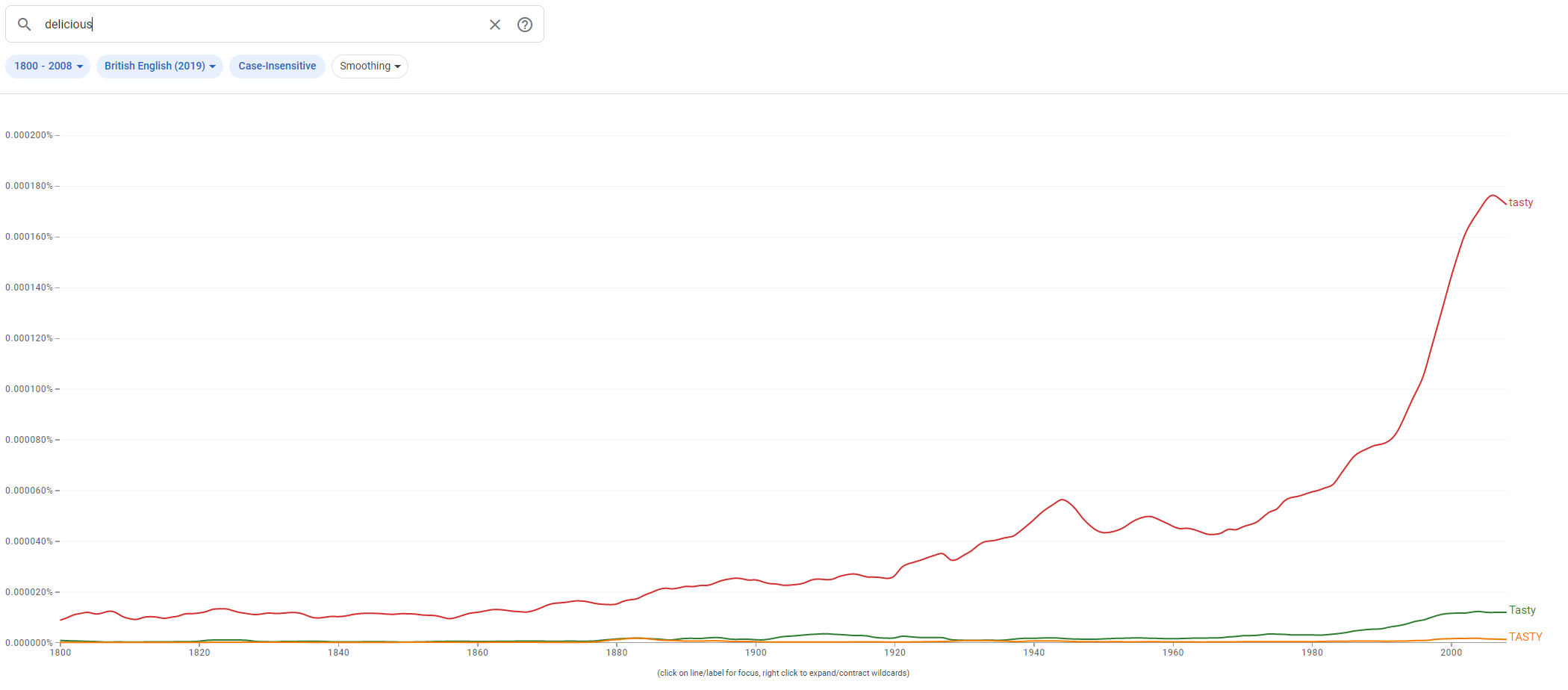


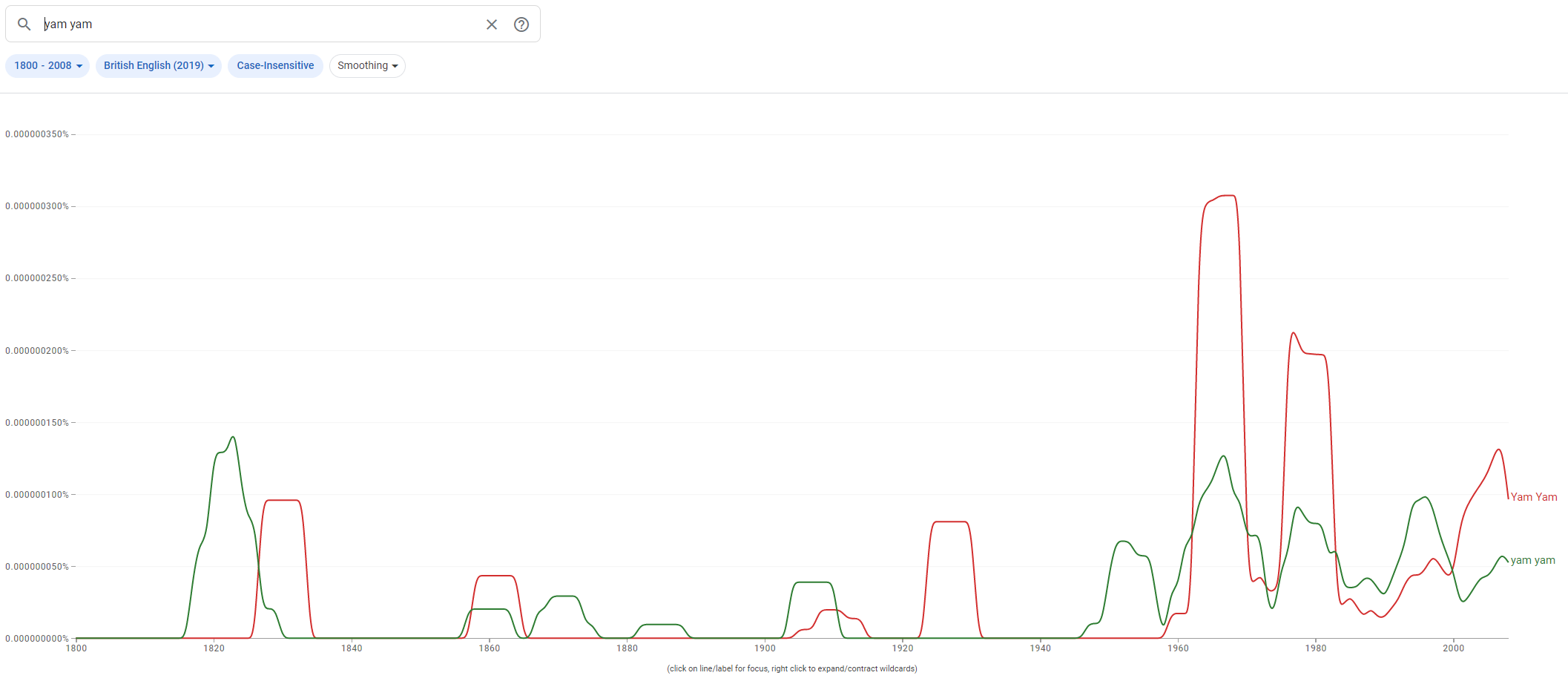




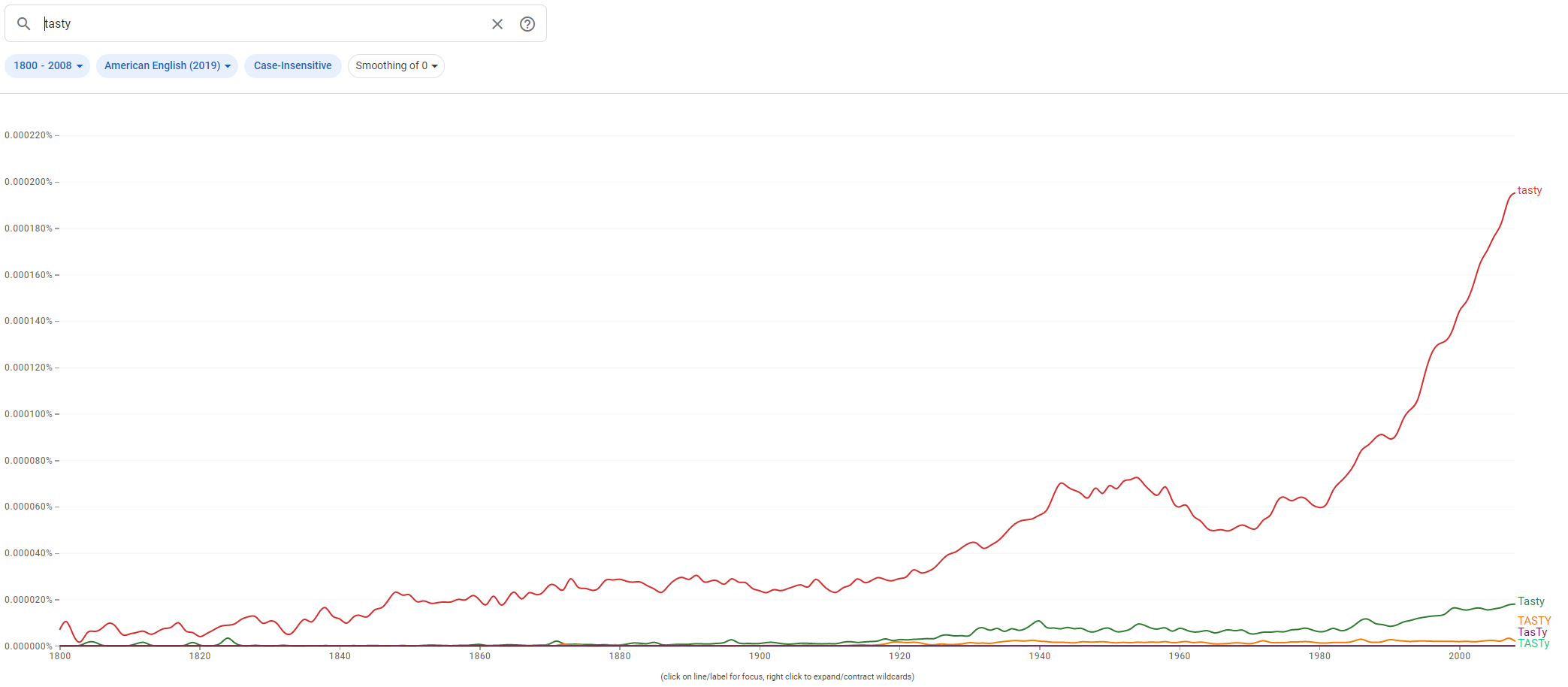
Примеры запросов tasty, delicious и yam yam соответственнов корпусе британского английского с 1800 по 2008 г. с параметром smoothing (сглаживание) 3:

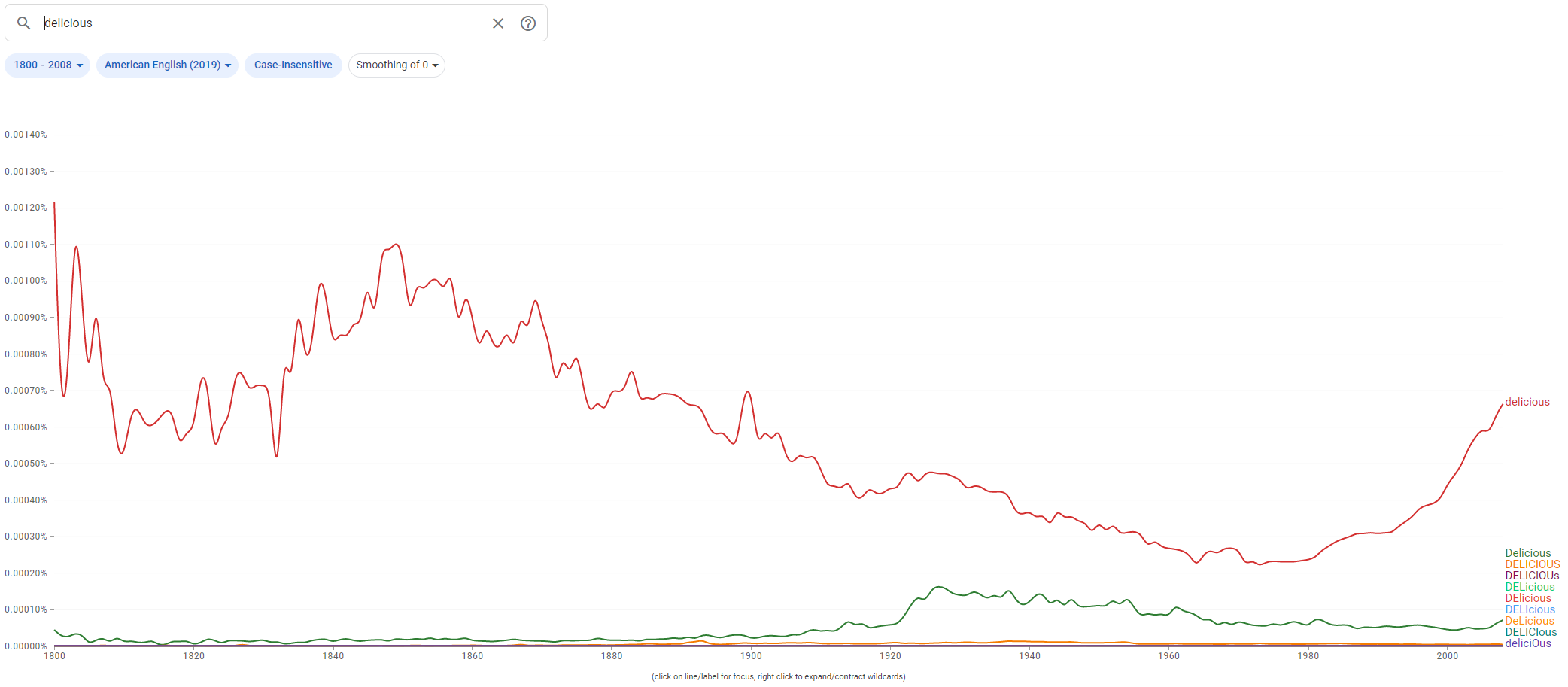


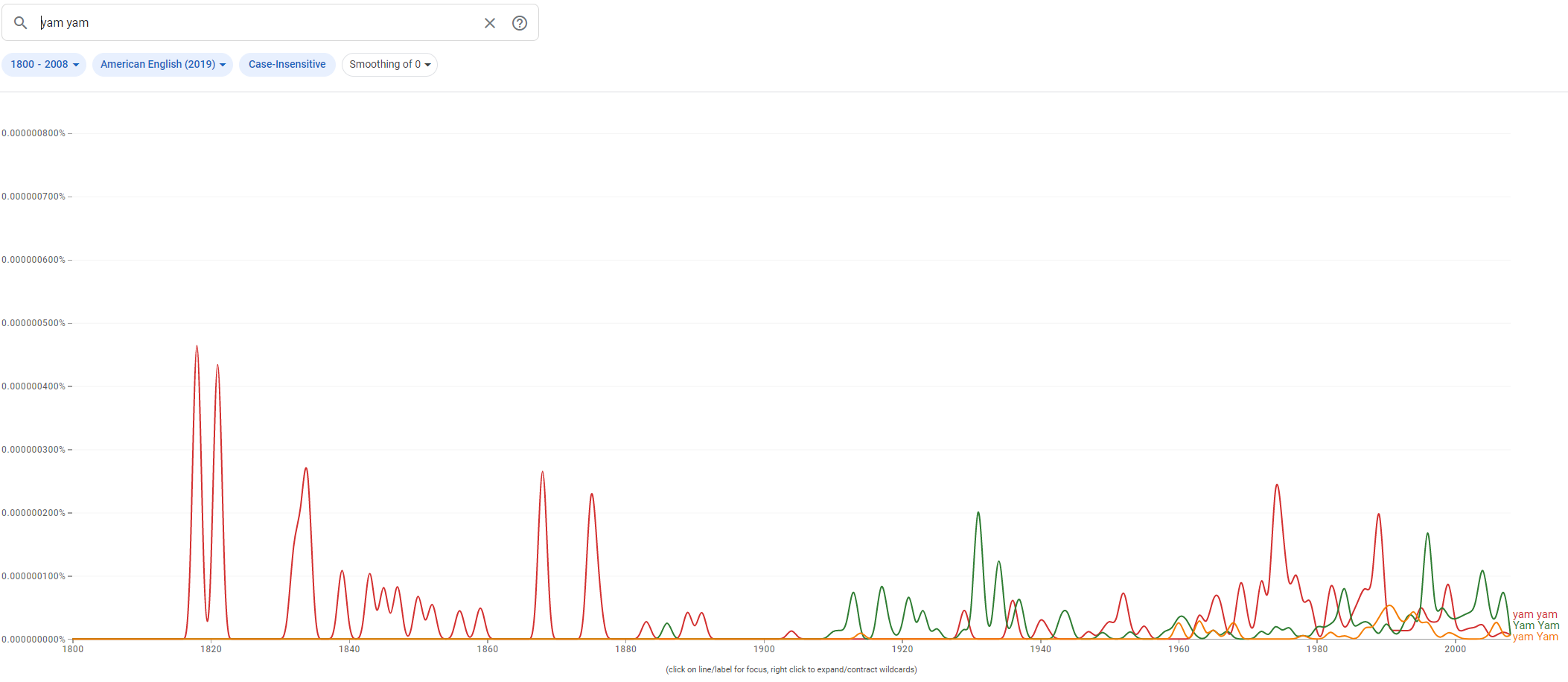




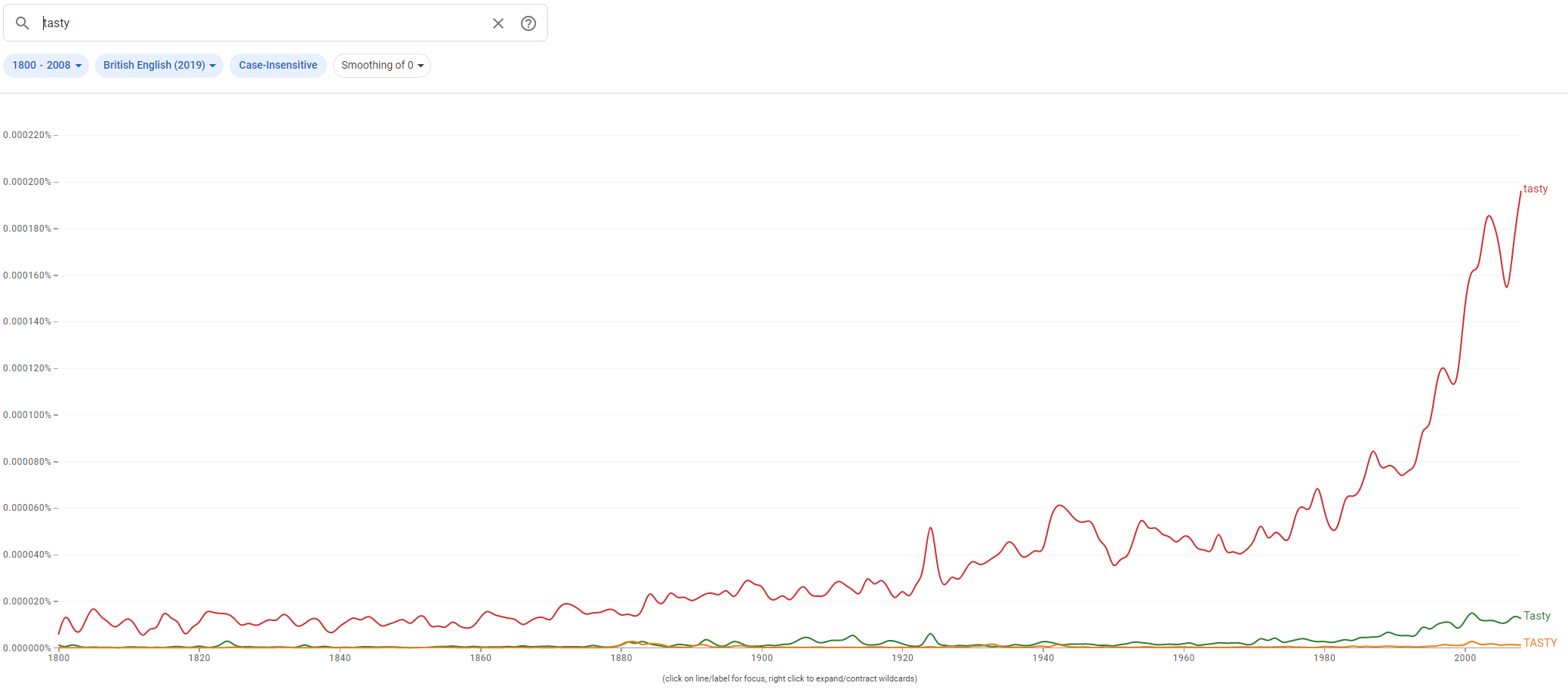
Примеры запросов tasty, delicious и yam yam соответственнов корпусе американского английского с 1800 по 2008 г. с параметром smoothing (сглаживание) 0:

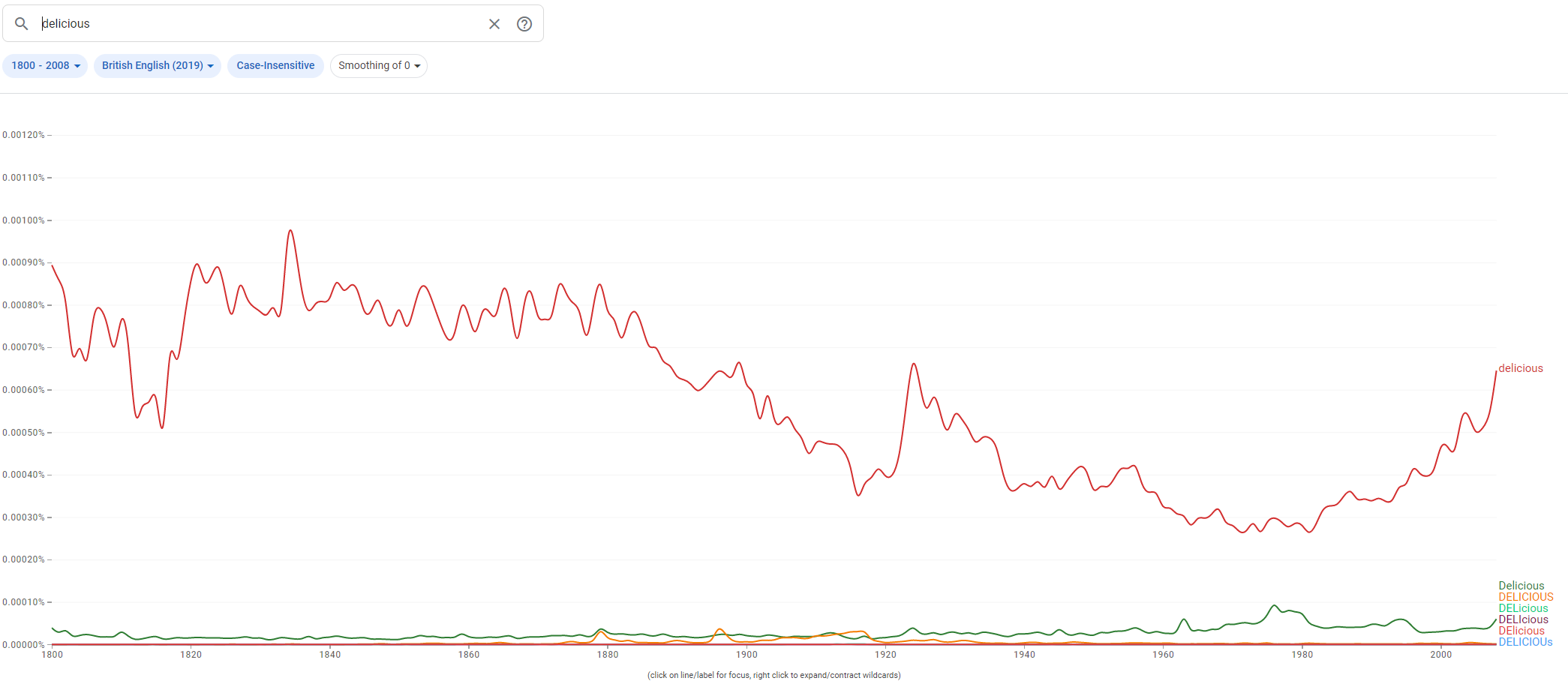


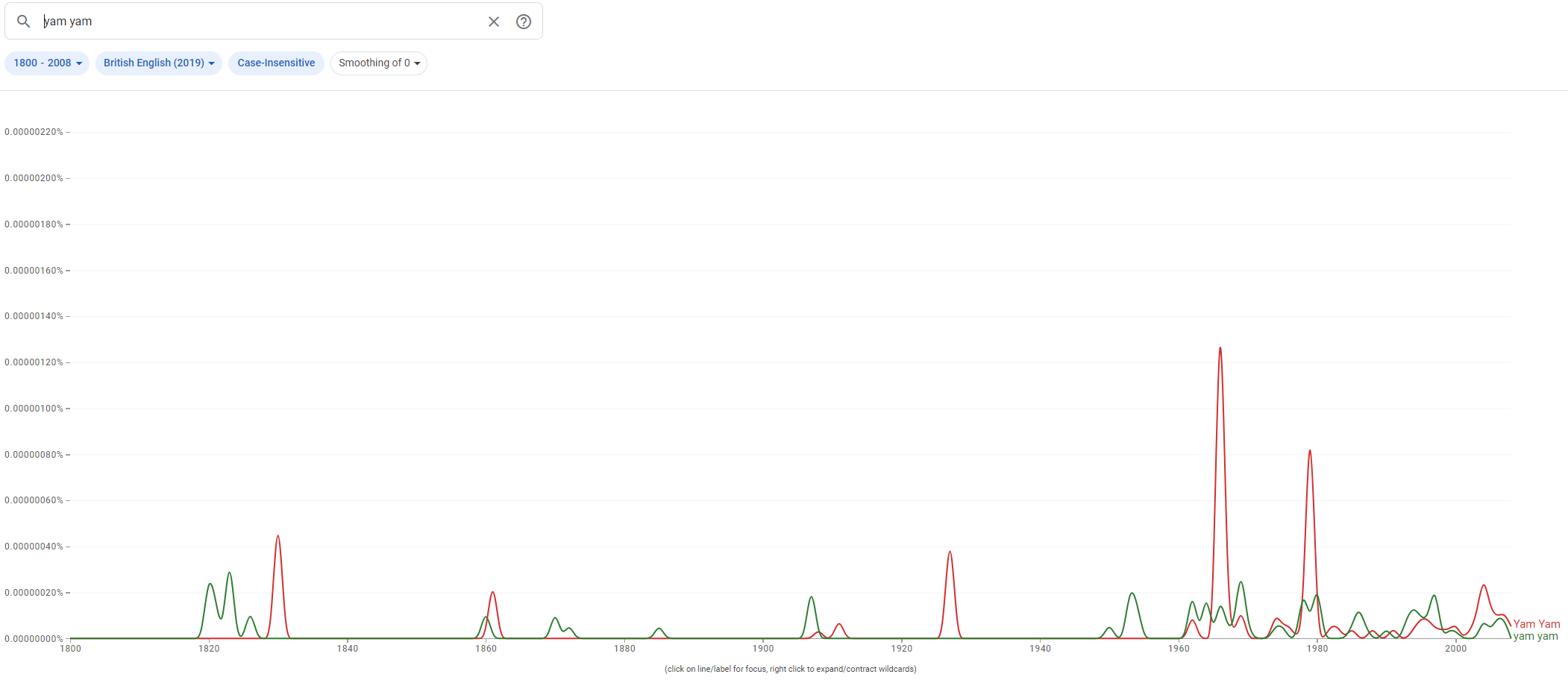




Примеры запросов tasty, delicious и yam yam соответственнов корпусе британского английского с 1800 по 2008 г. с параметром smoothing (сглаживание) 0:

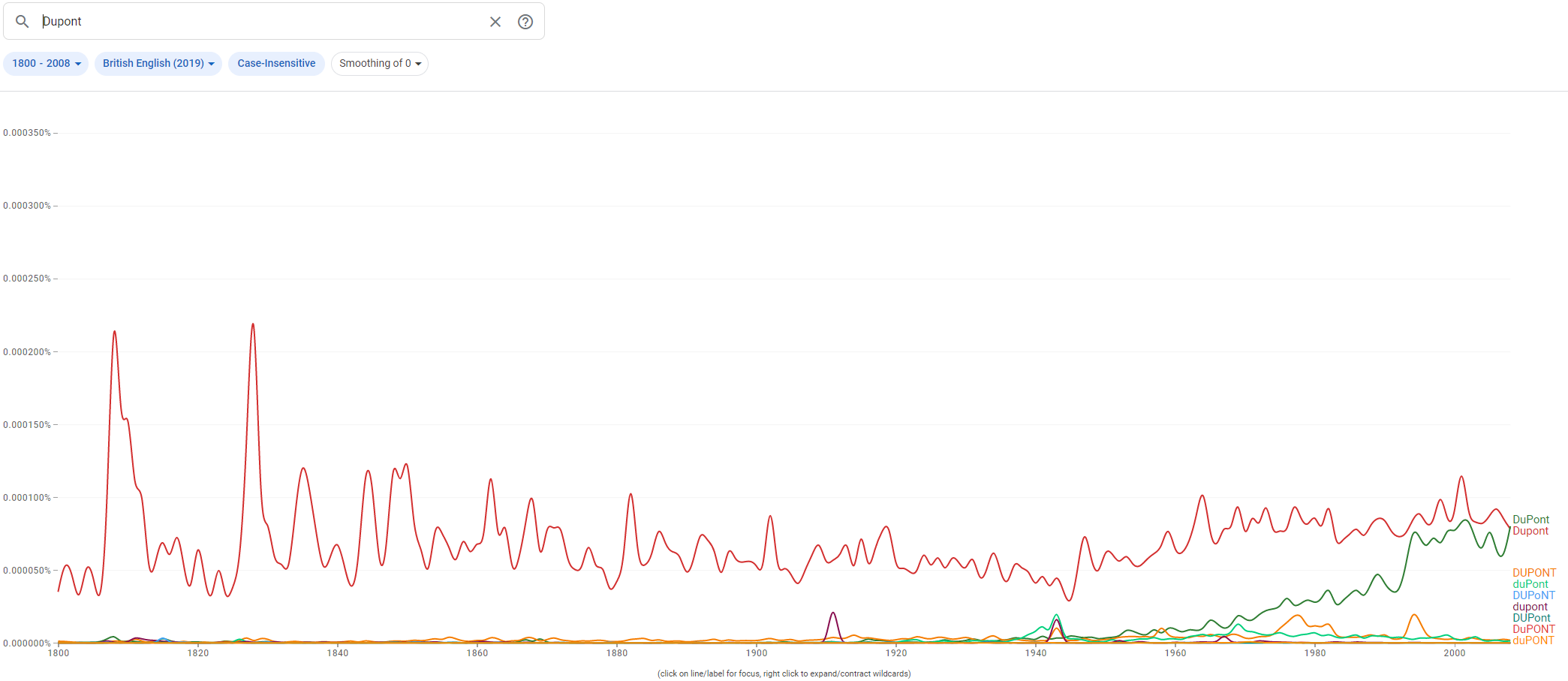




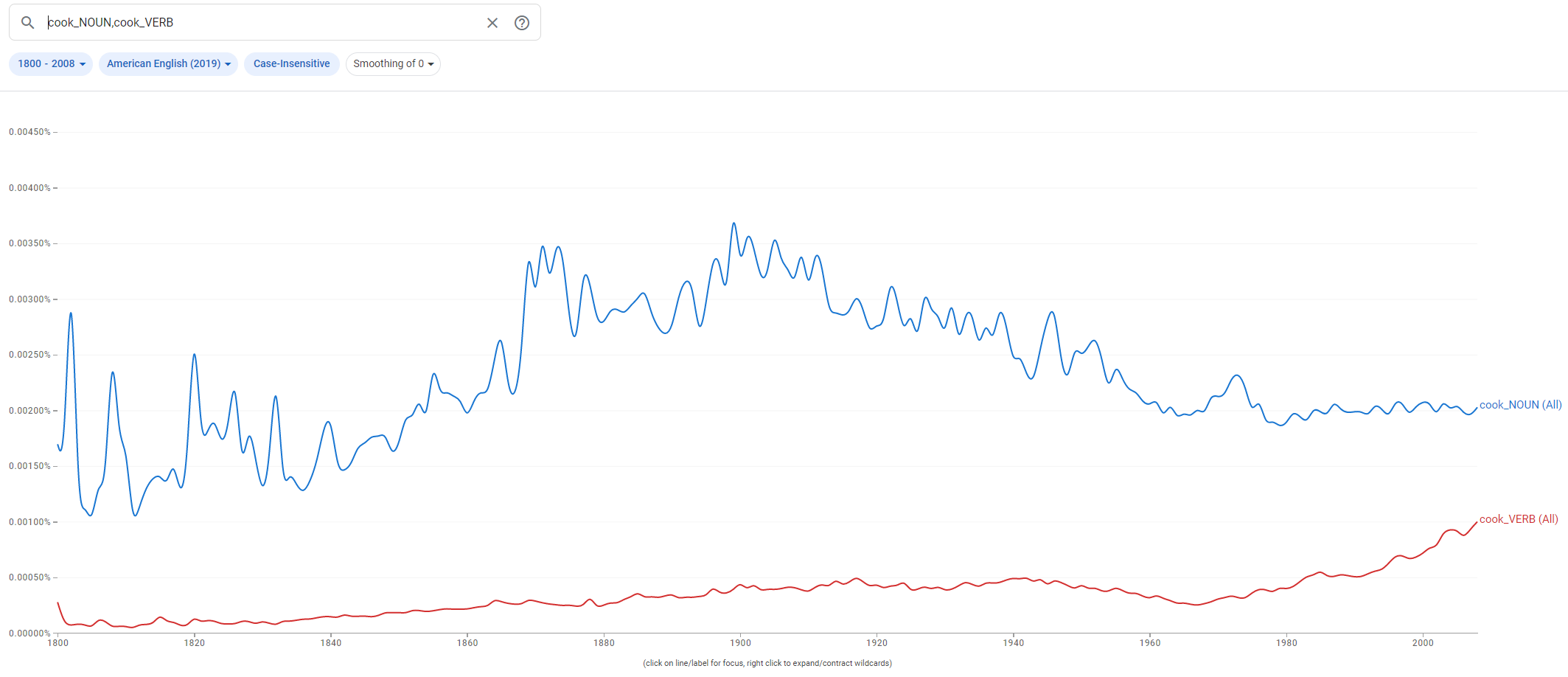


Смотря на графики (желательно со сглаживанием 0, так как на них лучше видно), можно сделать некоторые выводы. Все три слова из запросов обозначают примерно одно и то же, но имеют разную частотность. Видим, что “tasty” и “delicious” в корпусах британского и американского английского языков имеют схожие графики, стоит также отметить, что частота встречаемости слова “tasty” со временем росла и продолжает расти, тогда как слово “delicious”, напротив, имеет обратный график, то есть некогда, начиная с отметки 1800 года график начинался куда выше нынешних значений, хотя стоит сказать, что за последние 20-30 лет слово стали употреблять чаще. Отчасти такие обратные графики для слов, имеющих абсолютно одинаковые значения, можно объяснить так: слово “delicious” появилось раньше и соответственно использовалось чаще, в то время как слово “tasty” пришло ему на замену, как более упрощённая, менее утончённая и менее красочная форма, имеющая смысл “вкусный”. Со словом “yam yam” всё куда сложнее и непонятнее. Во-первых, можно увидеть, что в американском английском слово явно встречается чаще, чем в британском, но ни в одном, ни в другом корпусе оно нестабильно и непостоянно: весь график представлен только скачками, каких-либо плавных переходов у слова нет. Это может быть связано с его неоднозначностью и узконаправленностью, как я понимаю, в нашем контексте это сленговое выражение, обозначающее нечто вкусное, питательное и, скорее всего, неполезное. По графику видно, что слово хоть и скачками, но чаще употребляется в американском английском, нежели в британском.

Параметр smoothing отвечает за сглаживание всех “острых углов” графика, то есть убирает все мелкие колебания. Чем выше параметр, тем более ровный, плавный и прямой получается график. Высокие значения помогают легче воспринять и увидеть общую картину частотности определённого слова, уловить тенденцию и убрать возможно лишние шумы. В это же время более низкие значения, в частности 0, помогают увидеть полную картину встречаемости, что также может быть полезно.



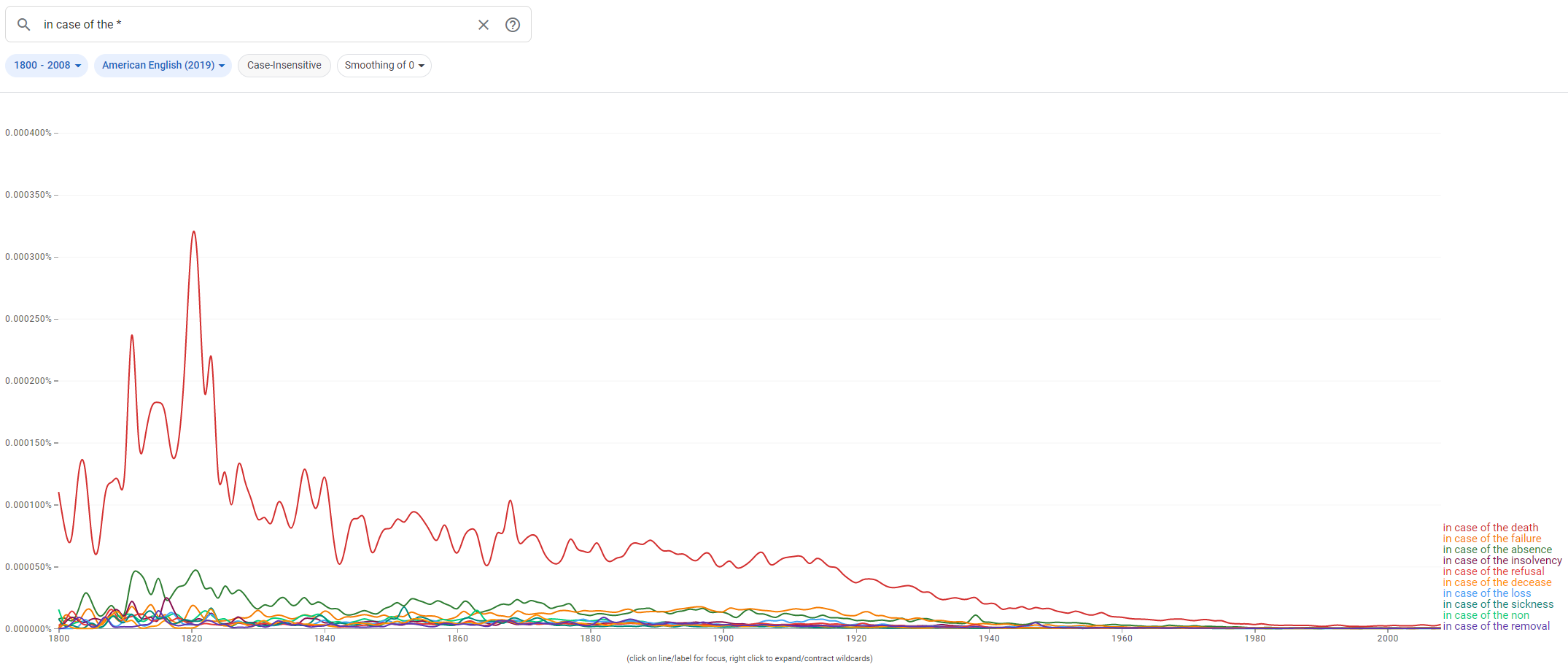
Из графика видно, что самая часто встречаемая форма слова Dupont – DuPont, что неудивительно, так как именно так называется американская компания.

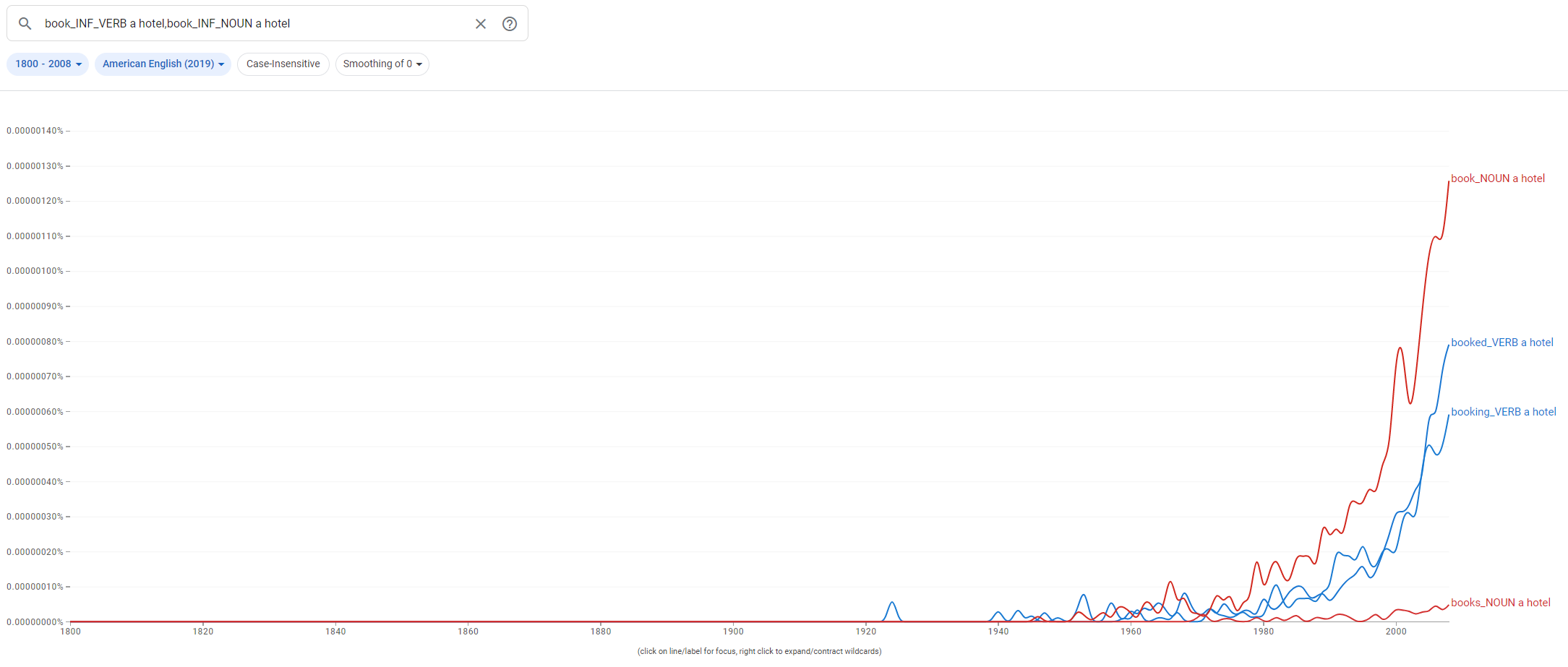


На примере корпуса американского английского можно увидеть, что слово “cook” чаще употребляется, как существительное.

Ниже на графике представлены самые частотные продолжения фразы “in case of the”. По нему легко можно заметить, что самое частое продолжение (с большим отрывом) – in case of the death (в случае смерти). Скорее всего это неудивительно, ведь раньше про смерть говорили чуть более смело: об этом спокойно писали в документах, могли это озвучить при заключении чего-либо. Сейчас же это чаще всего просто-напросто замалчивается, либо заменяется на менее пугающие и резкие высказывания.

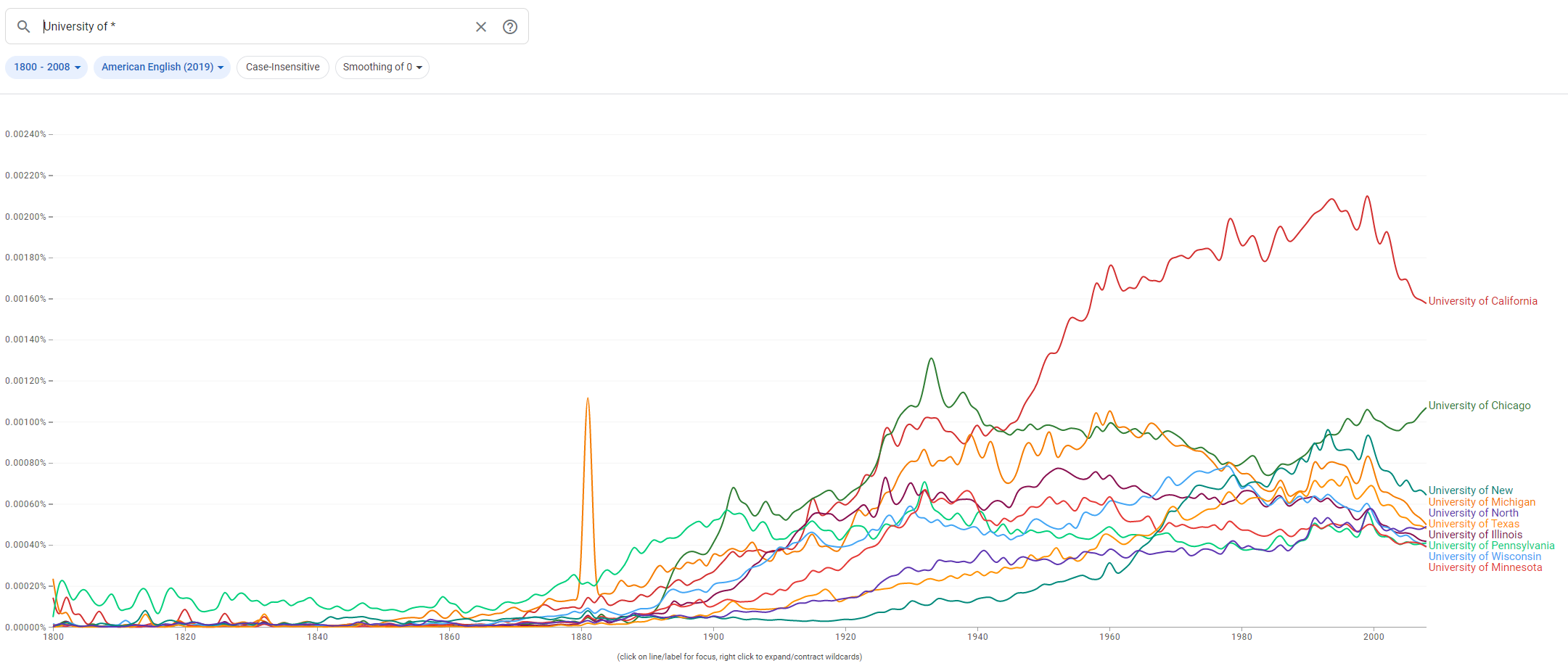
Также по графику можно увидеть, что в наше современное время нет какого-то определённого самого частотного продолжения фразы, все они сливаются в одну линию.



Странно, но график со словом “book” получился следующим: 

Я предполагал, что слово book в связке с “a hotel” однозначно будет чаще встречаться как глагол, ведь дословный перевод – забронировать отель, но результаты такие, какие есть. Либо я что-то сделал неправильно.

С институтами имеем следующее:



И сама конструкция:

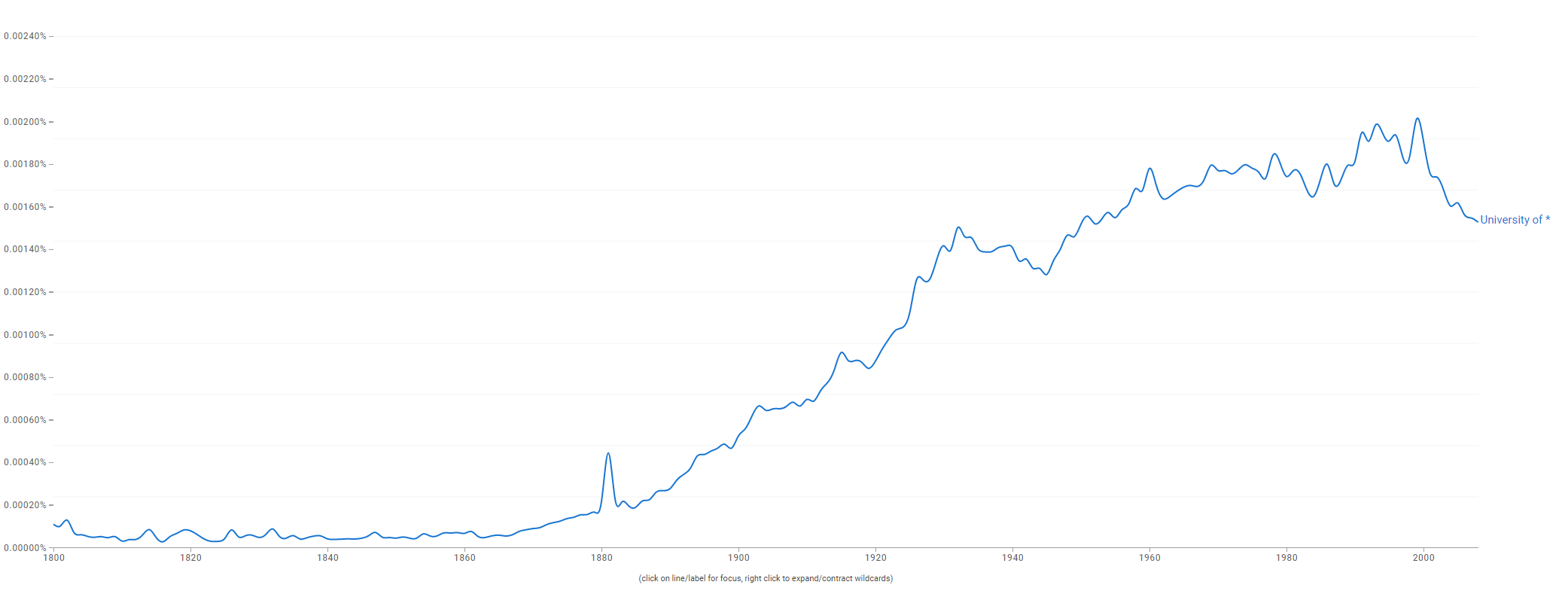
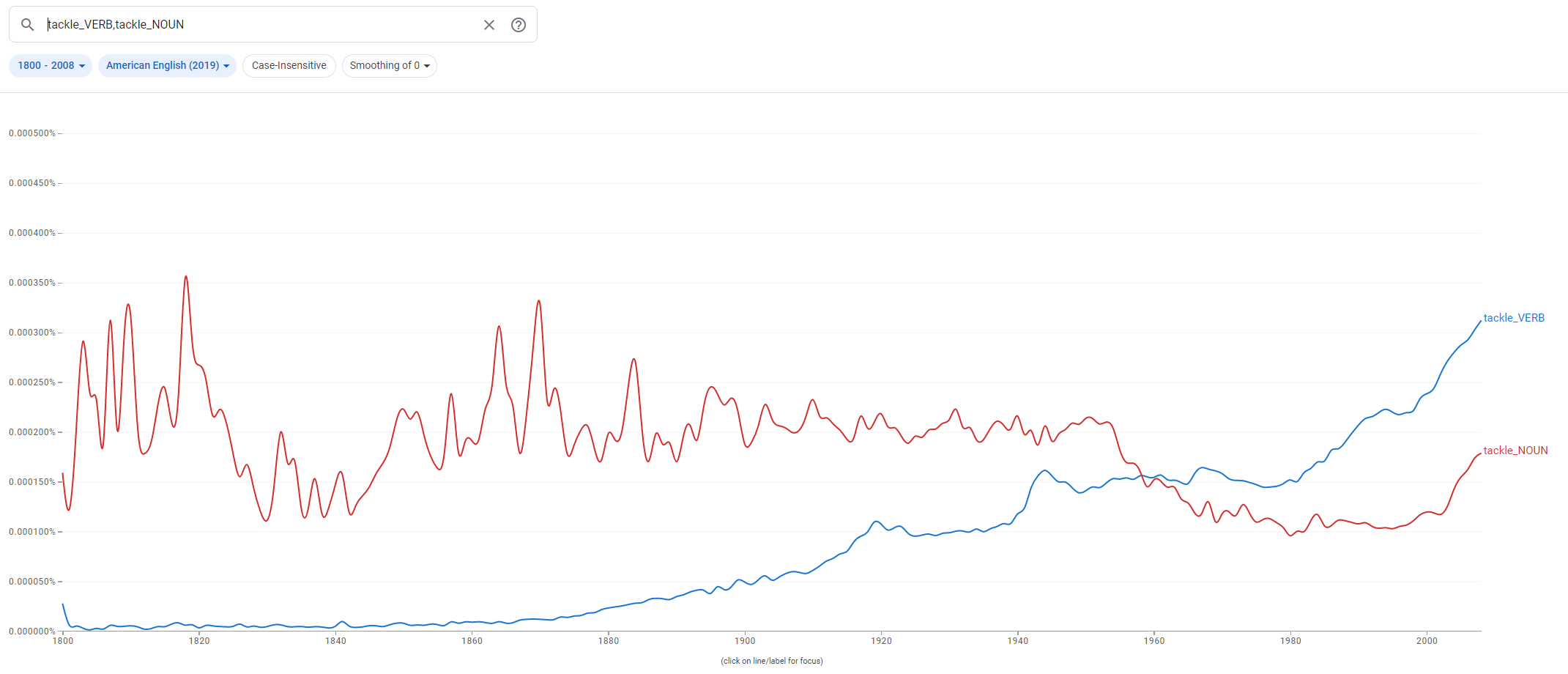
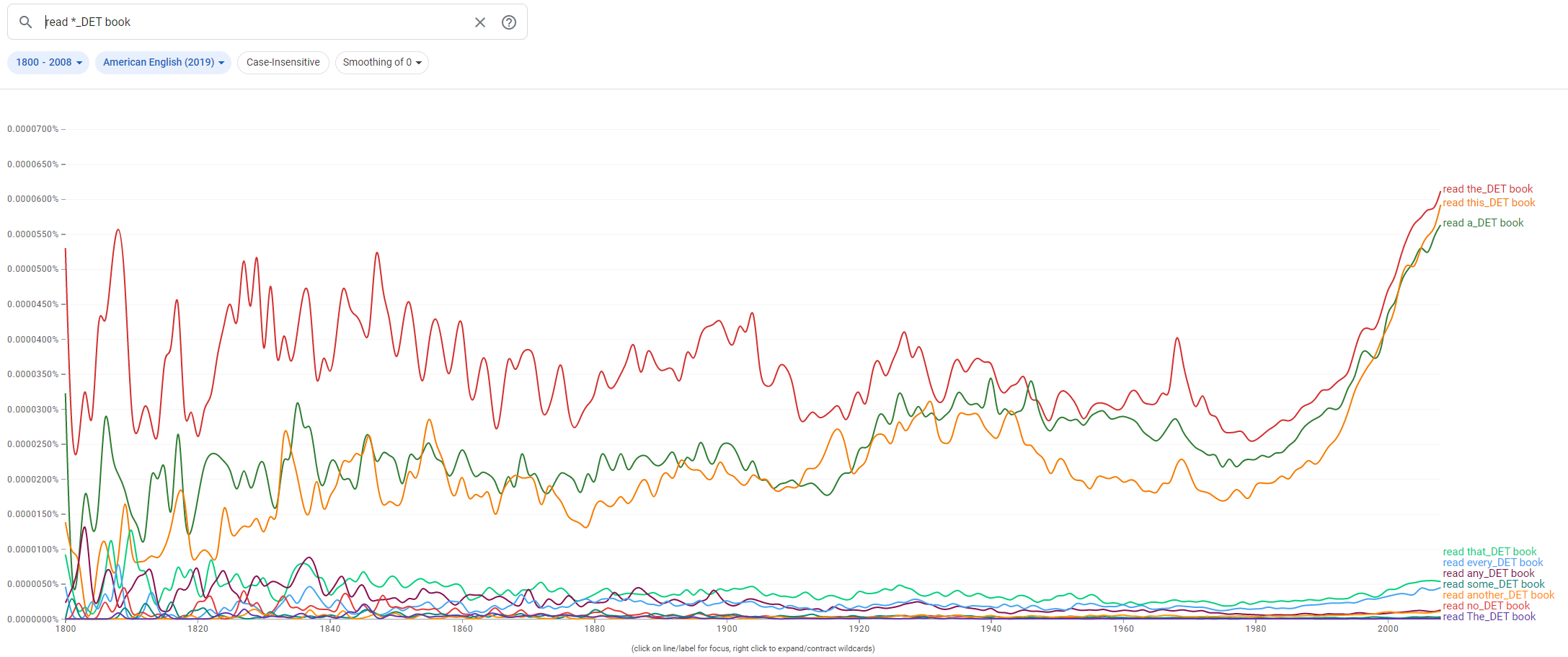


График встречаемости слова “tackle” как глагола и существительного:

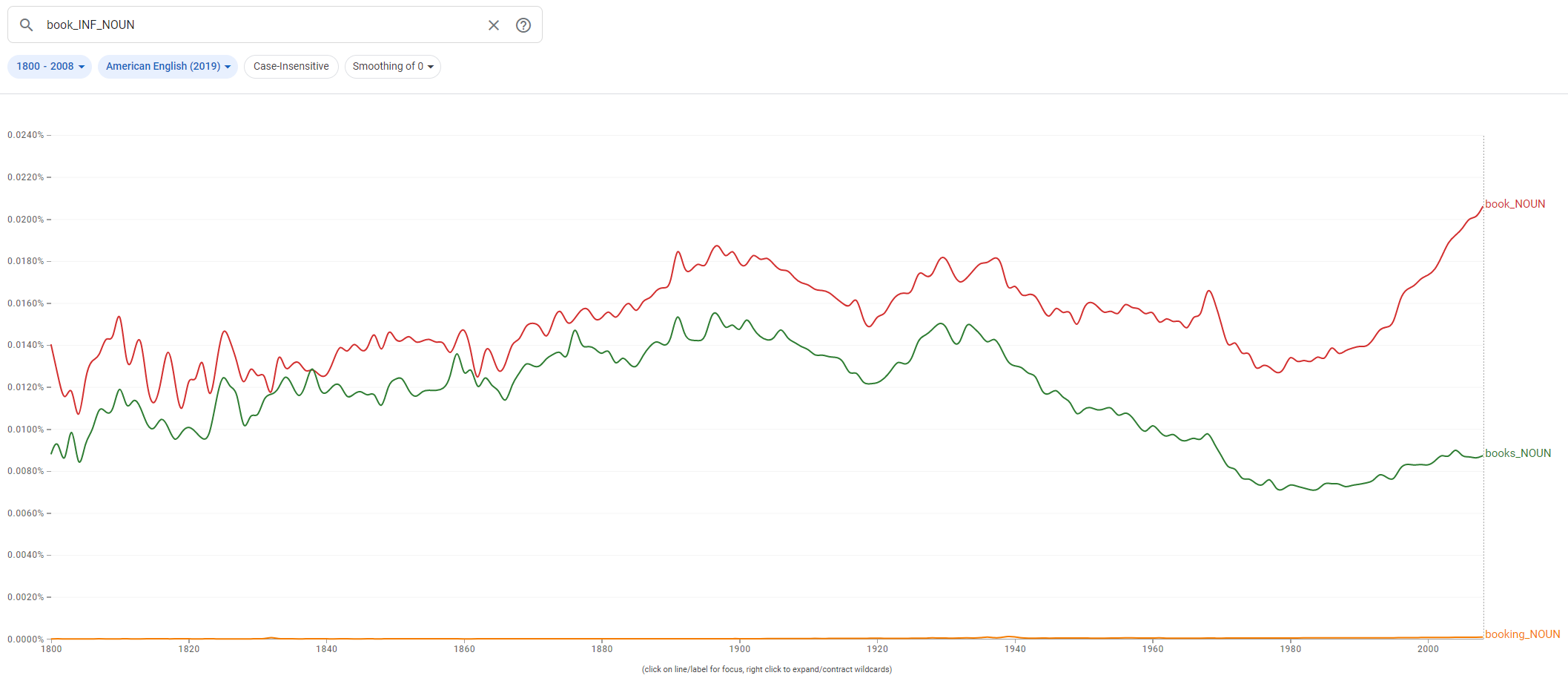


По нему видно, что большую часть времени версия существительного обгоняла по встречаемости, но начиная с 1958-го года глагольная форма обогнала существительное и до сих пор не сдаёт позиции.

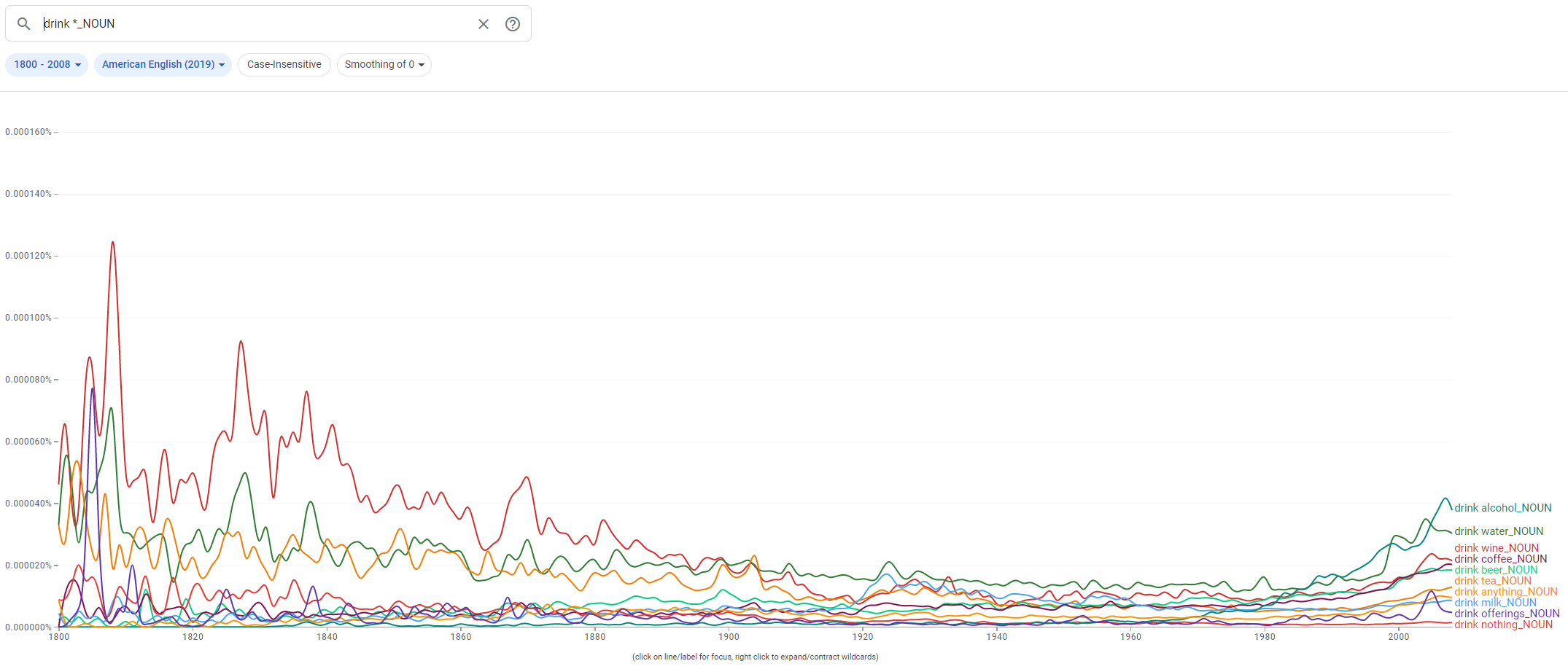
Имеем картину по артиклям во фразе read {article} book:



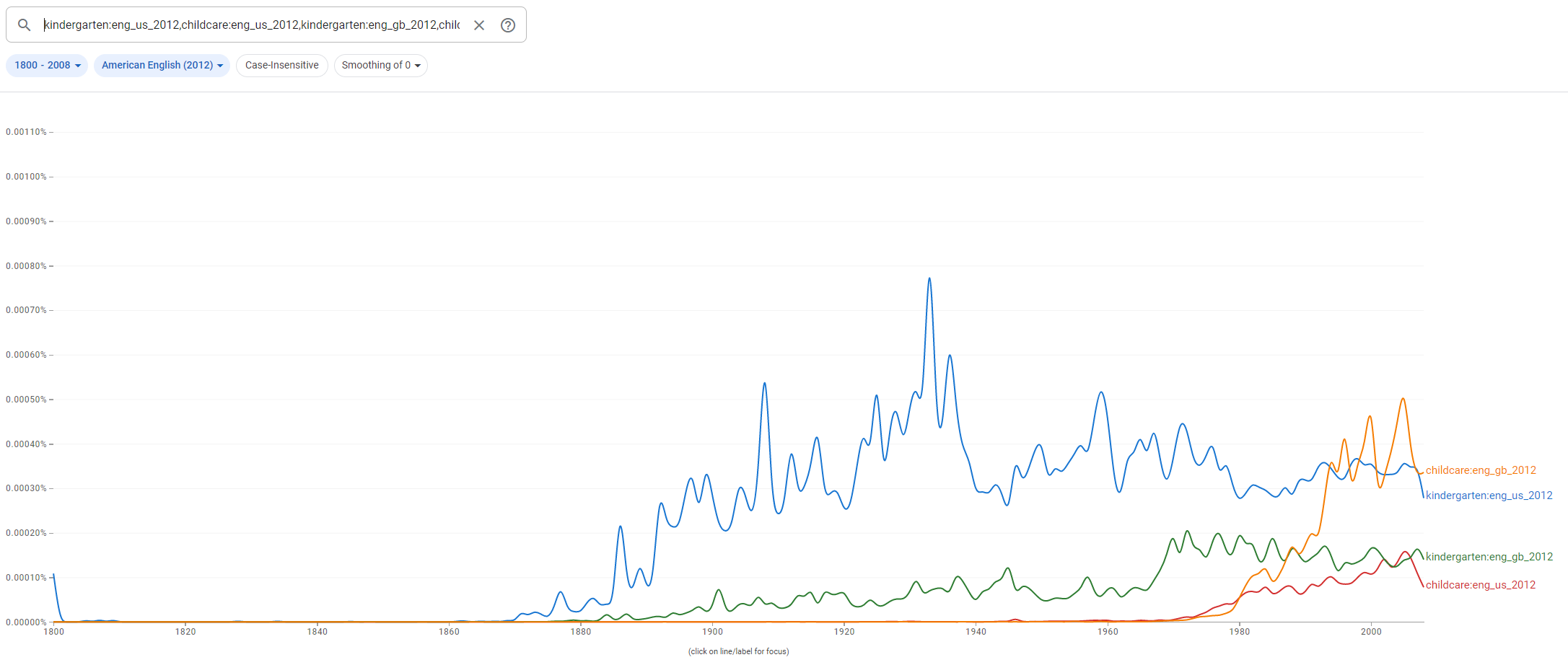
По графику окончаний слова “book”, когда после него идёт какое-либо существительное видно, что самые частые окончания этого слова – это -s, -ing или нулевое:



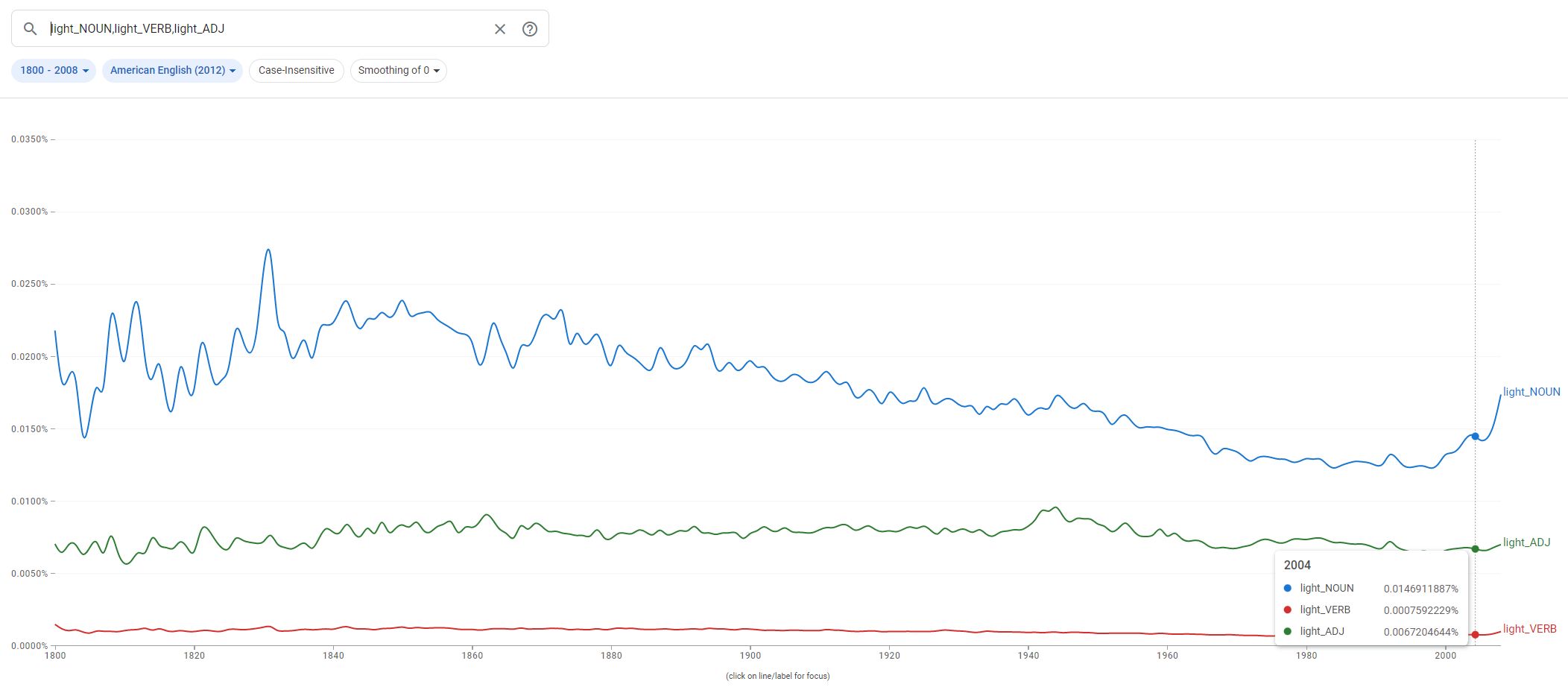
Чаще всего глаголом “drink” управляются такие существительные: алкоголь, вода, вино и т.д.:

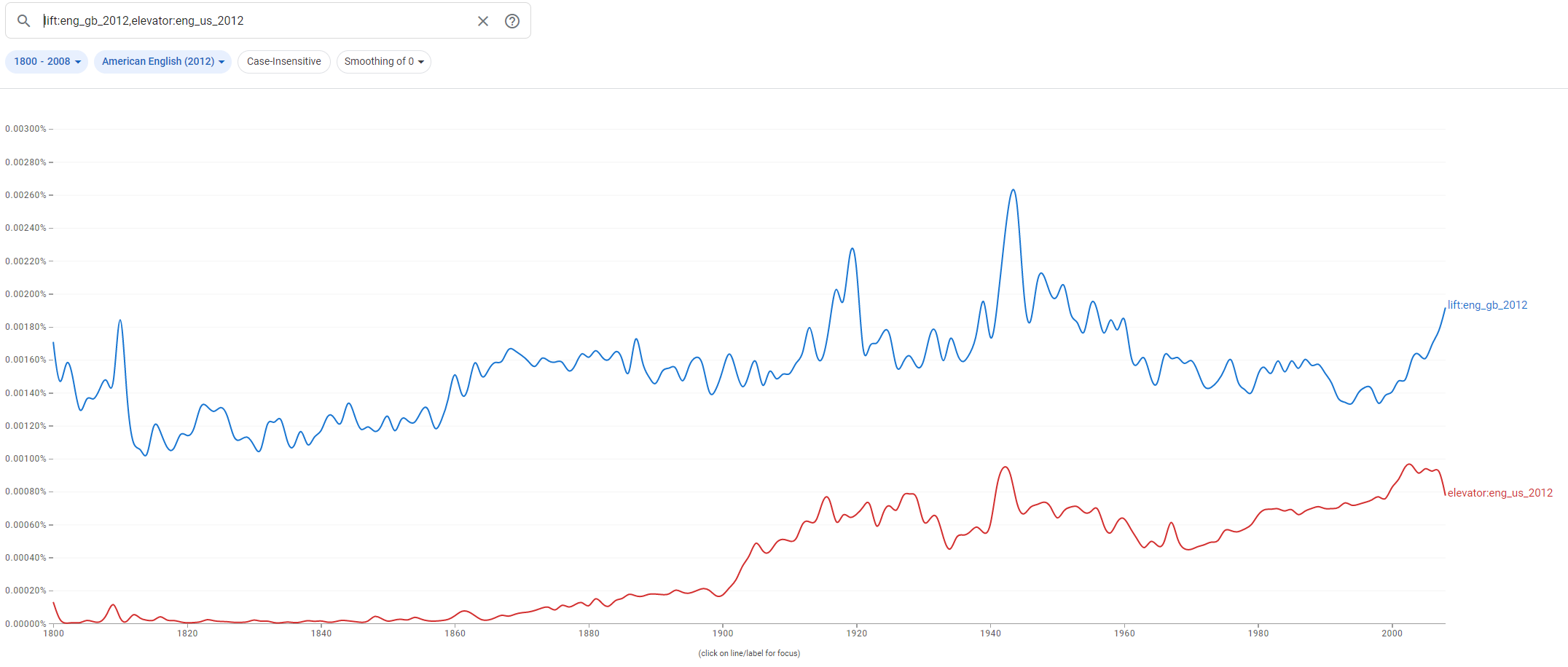


Сравнение частоты употребления слов kindergarten и childcare в корпусах британского и американского английского (версия 2012 г.):



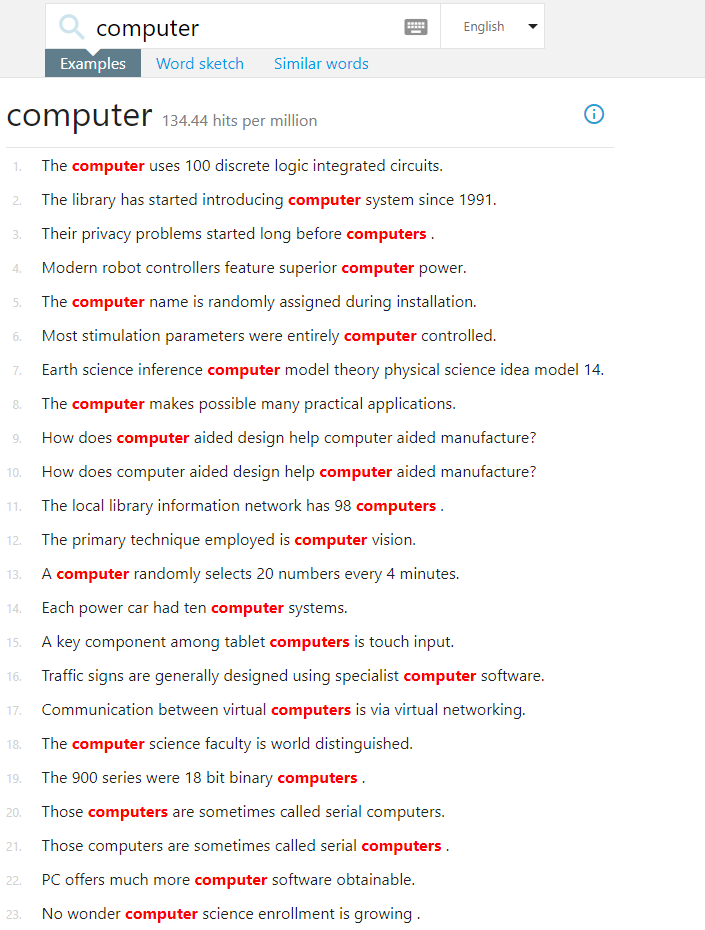
Примером слова, которое одновременно может быть представлено в трёх различных частях речи, может являться слово “light”, которое может быть как существительным, так прилагательным или глаголом:



Пример пары синонимов, характерных для определённого корпуса языка:  


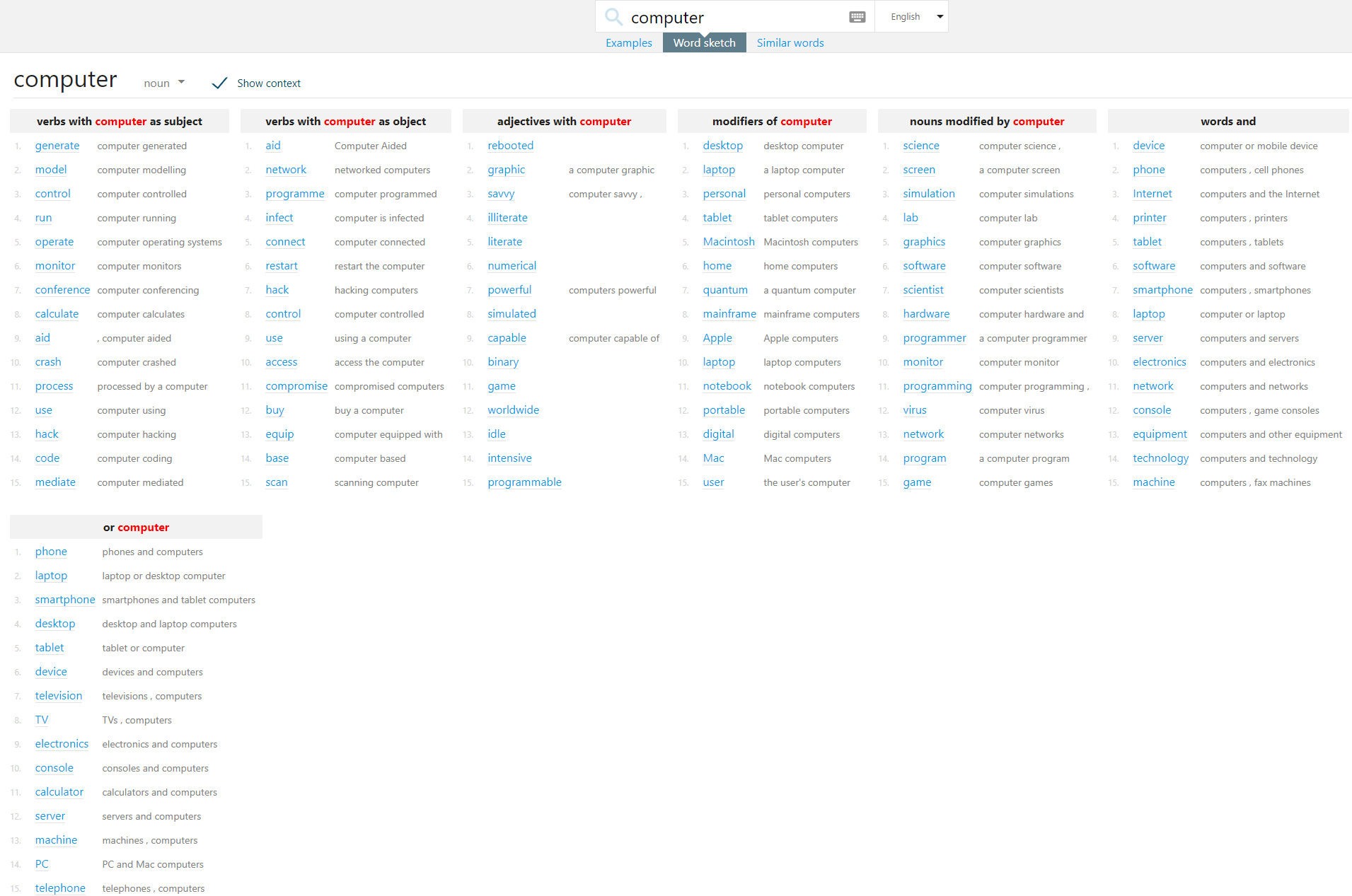
Мой опыт (в виде скриншотов) знакомства со SkeLL.

Слово “компьютер” в примерах:



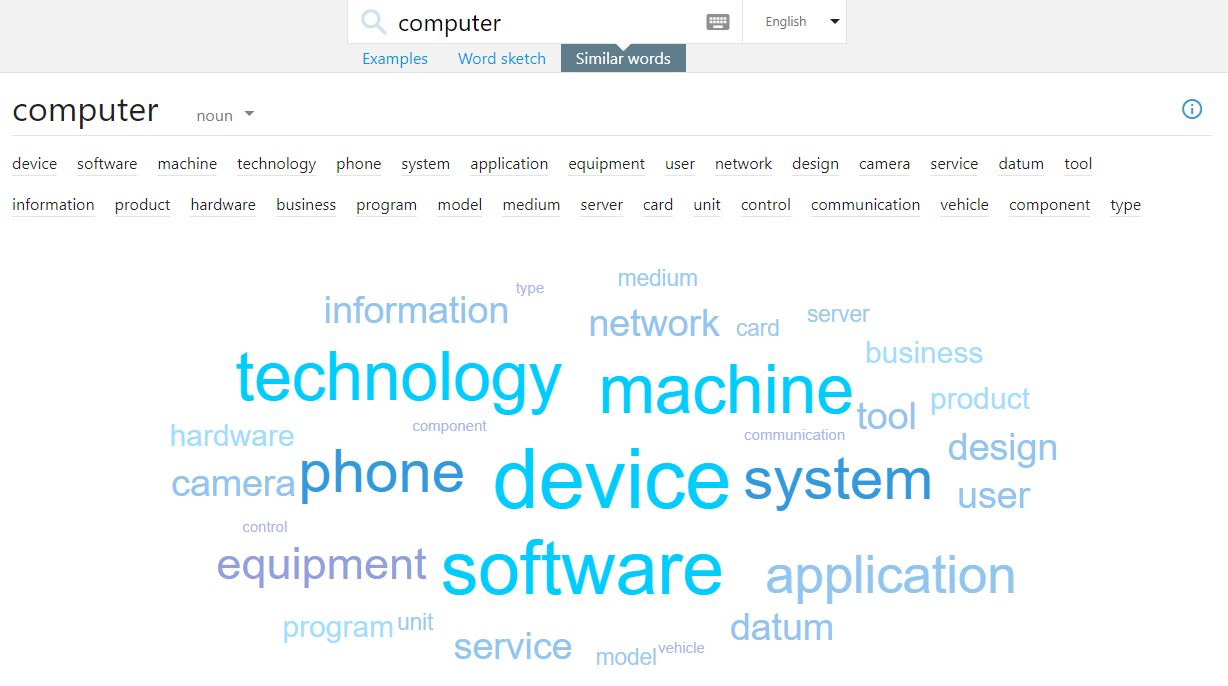
Данная функция помогает понять контекст и верное использование слова в предложении. Примеры могут быть полезны при изучении нового языка или, к примеру, при знакомстве с новыми узконаправленными словами или же при лингвистических исследованиях.

Слово “компьютер” в сочетаемости слов:



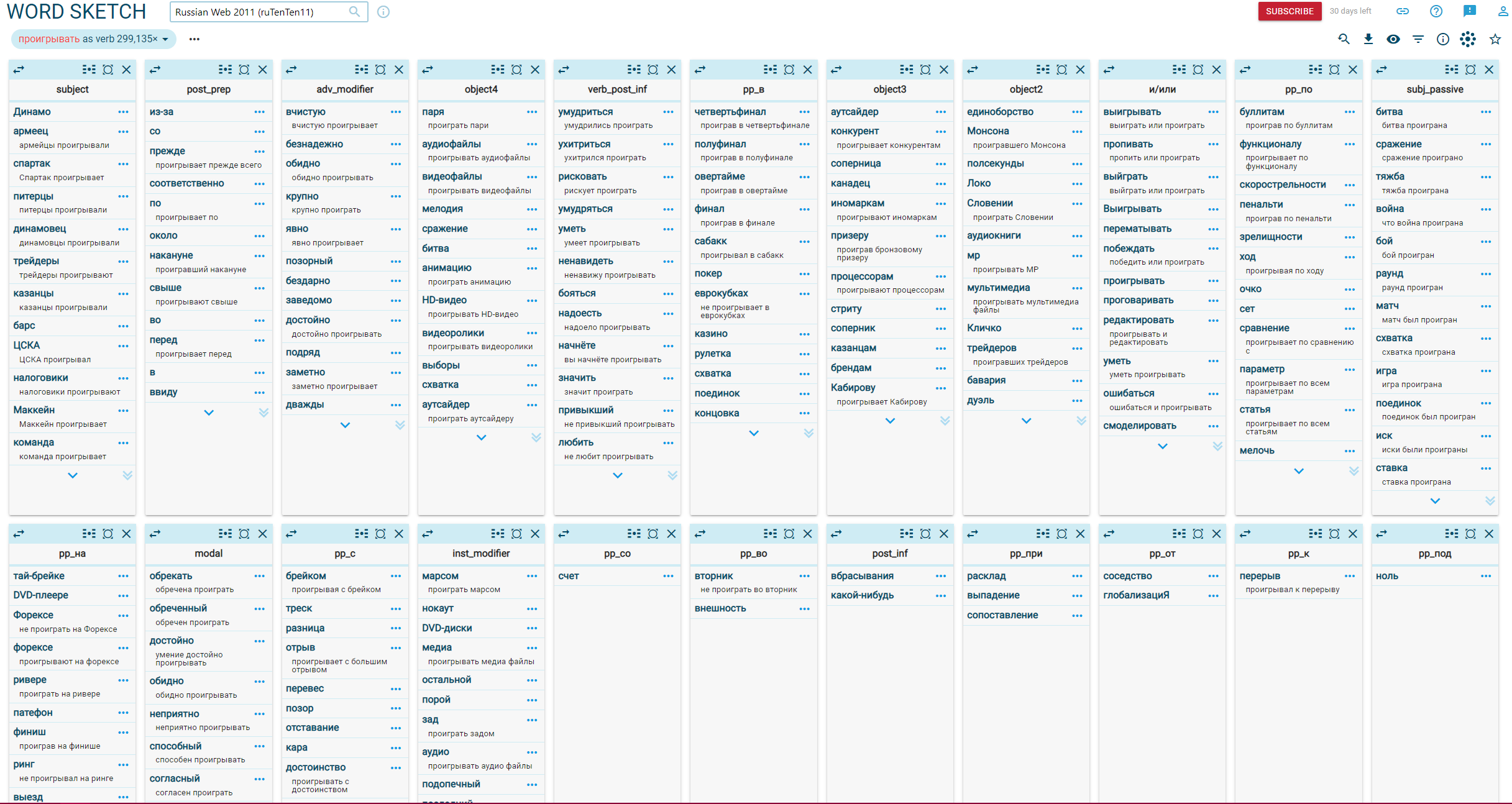
Данная же функция помогает понять то, с какими словами сочетается введённый термин, позволяет определить типичные лексические сочетания и идиомы, которые используются с введённым словом. Может также помочь в изучении нового языка, также просто при переводе позволяет находить переводчикам наиболее подходящие сочетания слов, чтобы они звучали естественно.

Слово “компьютер” в похожих словах:



Похожие слова позволяют пользователю найти синонимы, антонимы, иногда однокоренные слова и родственные понятия, которые связаны с исходным словом по смыслу. Одной из ключевых сфер применения данной функции помимо лингвистических исследований, конечно же, является расширение словарного запаса, что в свою очередь может пригодиться, как лингвистам, переводчикам, просто студентам, так и случайному пользователю.

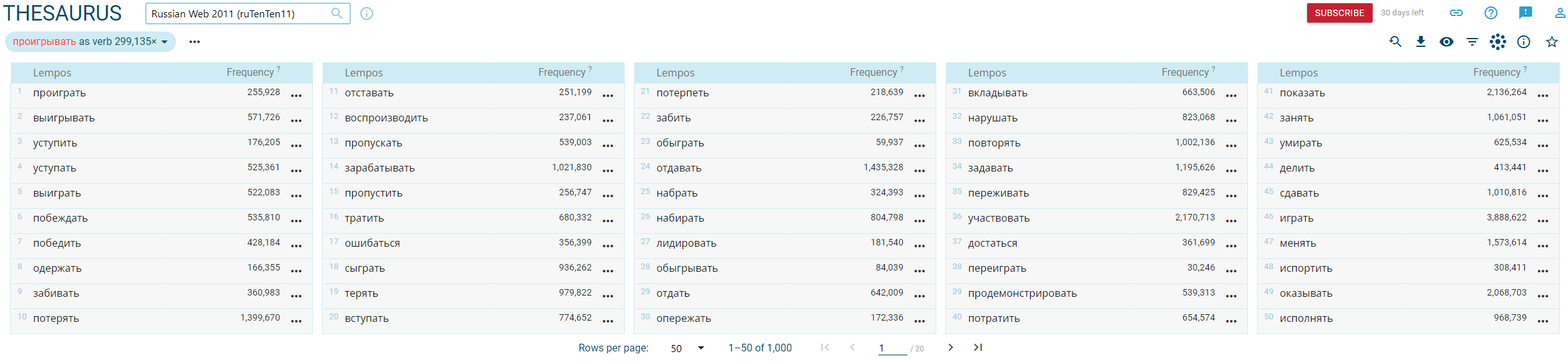
Пример работы с Word Sketch:



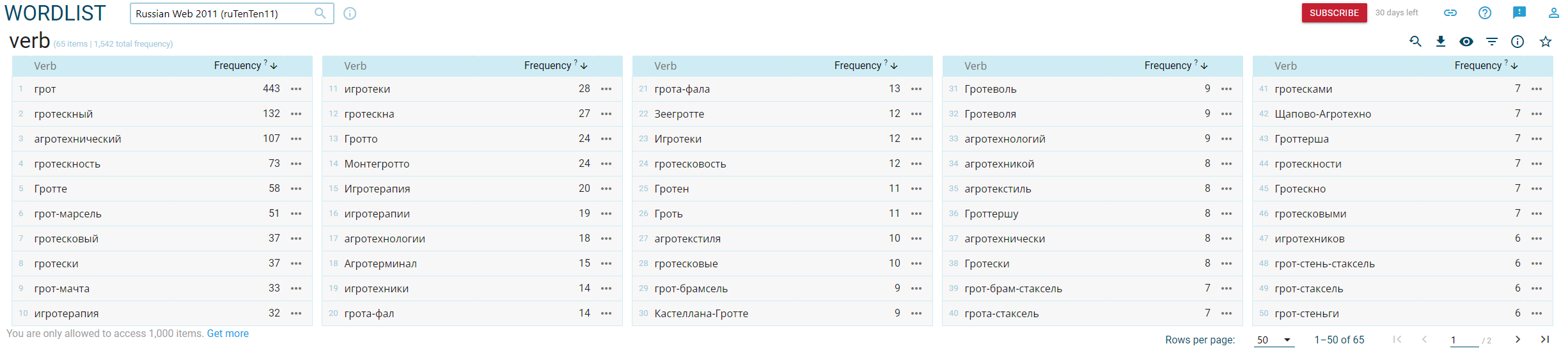
Пример работы с Word Sketch Difference:



Пример работы с функцией Trends в корпусе британского английского языка:  

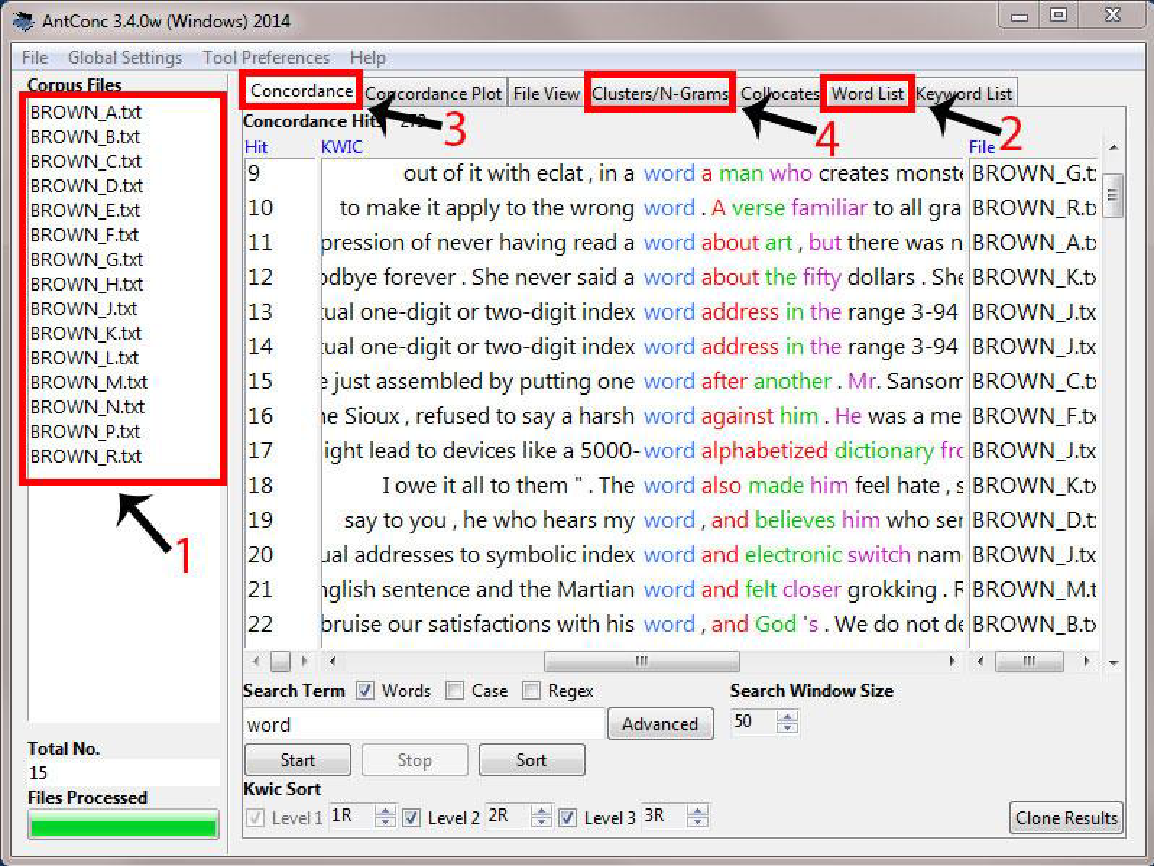

Пример работы с тезаурусом:  


Пример работы с Wordlist:



Можно подвести следующий небольшой итог:

* Word Sketch (лексический образ) в Sketch Engine предоставляет подробные описания слов и их употребление в контексте. Он показывает типичные словосочетания, грамматические конструкции и семантические роли, связанные с выбранным словом. Word Sketch помогает лучше понять употребление слова в реальных текстах и обнаруживать его лингвистические особенности
* Word Sketch Difference (разница лексических образов) позволяет сравнивать лексические образы двух слов и выявлять семантические, грамматические и стилистические различия между ними. Это полезный инструмент для сопоставления слов и анализа их семантических оттенков и употребления
* Trends (Тренды) позволяет анализировать изменения в частоте употребления слов и фраз в различных текстовых корпусах со временем. Она показывает графическое представление данных, которое помогает определить тенденции, популярность и сезонные колебания в использовании определенных лексических единиц
* Thesaurus предоставляет синонимы, антонимы и связанные слова для заданного слова или фразы. Это полезный инструмент для расширения словарного запаса, поиска альтернативных выражений и изучения семантических связей между словами
* Wordlist позволяет создавать и работать с заданными списками слов, основанными на заданных параметрах и фильтрах. Это полезный инструмент для составления словарей, изучения словосочетаний или отслеживания частотности и употребления определенных слов



1. Вкладка “Corpus Files” предназначена для загрузки и управления корпусными файлами. Вы можете добавлять новые файлы в ваш корпус или удалять уже существующие файлы. Вы также можете просмотреть список загруженных файлов, их имена, пути к файлам и другую информацию о каждом файле в корпусе.
2. Вкладка “Word List” позволяет создать список слов, найденных в выбранном корпусе или текстовом файле. Здесь можно увидеть частотность слов, сортировать их по алфавиту или частоте, а также применять различные фильтры и настройки
3. Вкладка “Concordance” позволяет просматривать контекстные строки, содержащие искомые слова. Здесь можно увидеть слова, окружающие искомое слово, и изучить его употребление в различных контекстах
4. Вкладка “Cluster/N-Grams” позволяет анализировать частотность групп слов (кластеров) и n-грамм в тексте. Здесь можно найти типичные словосочетания, фразы и семантические шаблоны.

Немного о программе AntConc:

Инструмент Concordance позволяет просматривать контексты, в которых встречается искомое слово или фраза. Пользователь может увидеть предложения, в которых слово используется, и изучить его использование в различных контекстах. Это полезно для изучения семантики, коллокаций и структуры предложений.

Функция Clusters позволяет анализировать частотность групп слов (кластеров) и n-грамм в тексте. Он помогает идентифицировать типичные словосочетания, фразы и семантические шаблоны, которые встречаются в тексте с определенной частотой.

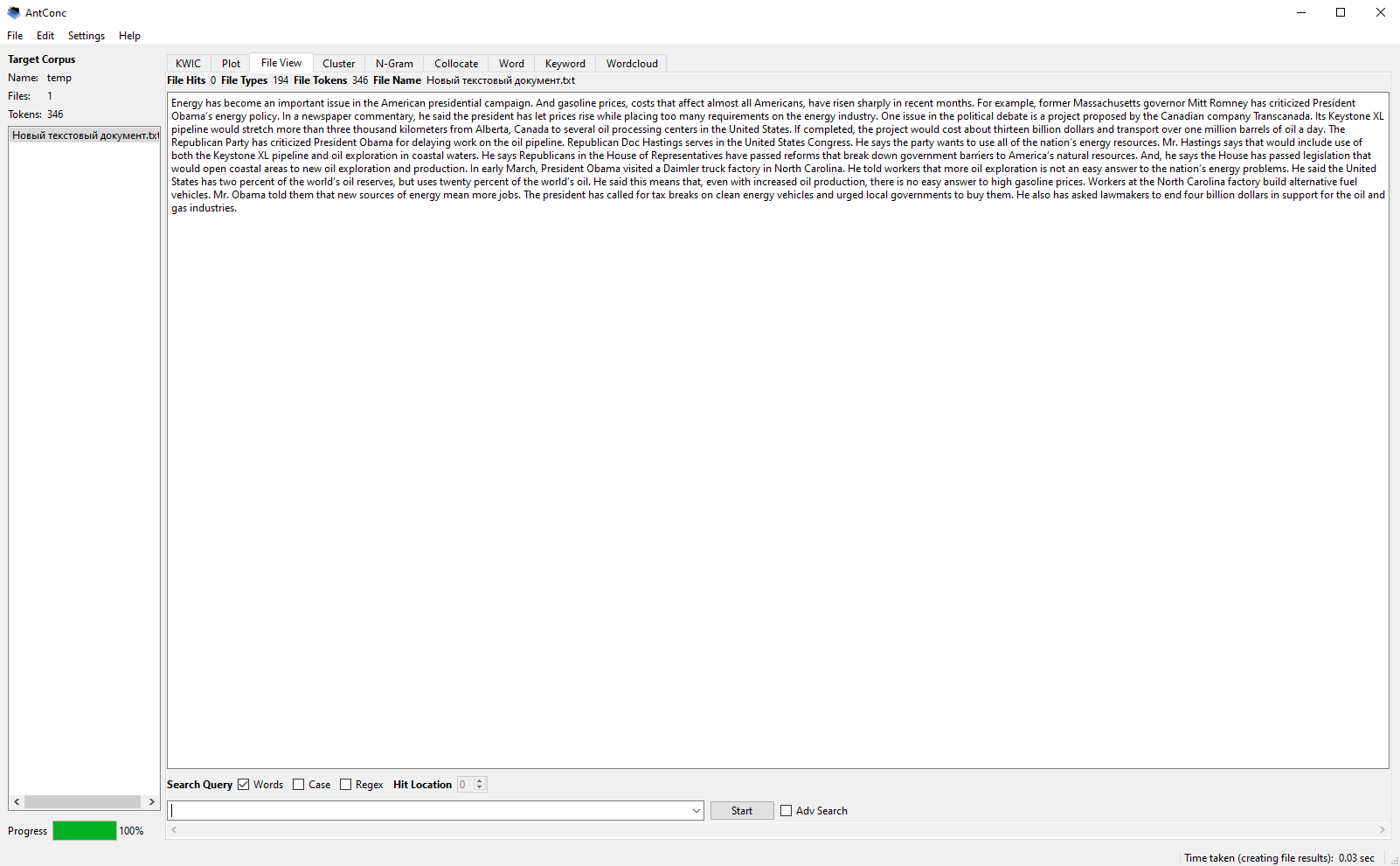
Word List позволяет создать список слов, найденных в тексте, и отобразить их частотность. Пользователь может отсортировать слова по частотности и изучать лексические особенности текста на основе использования разных слов.

Инструмент Key Words позволяет определить ключевые слова и термины, которые отличаются по своей частотности и употреблению в заданном тексте или корпусе по сравнению с другими корпусами. Он помогает выделить особенности и специализированные термины в тексте.

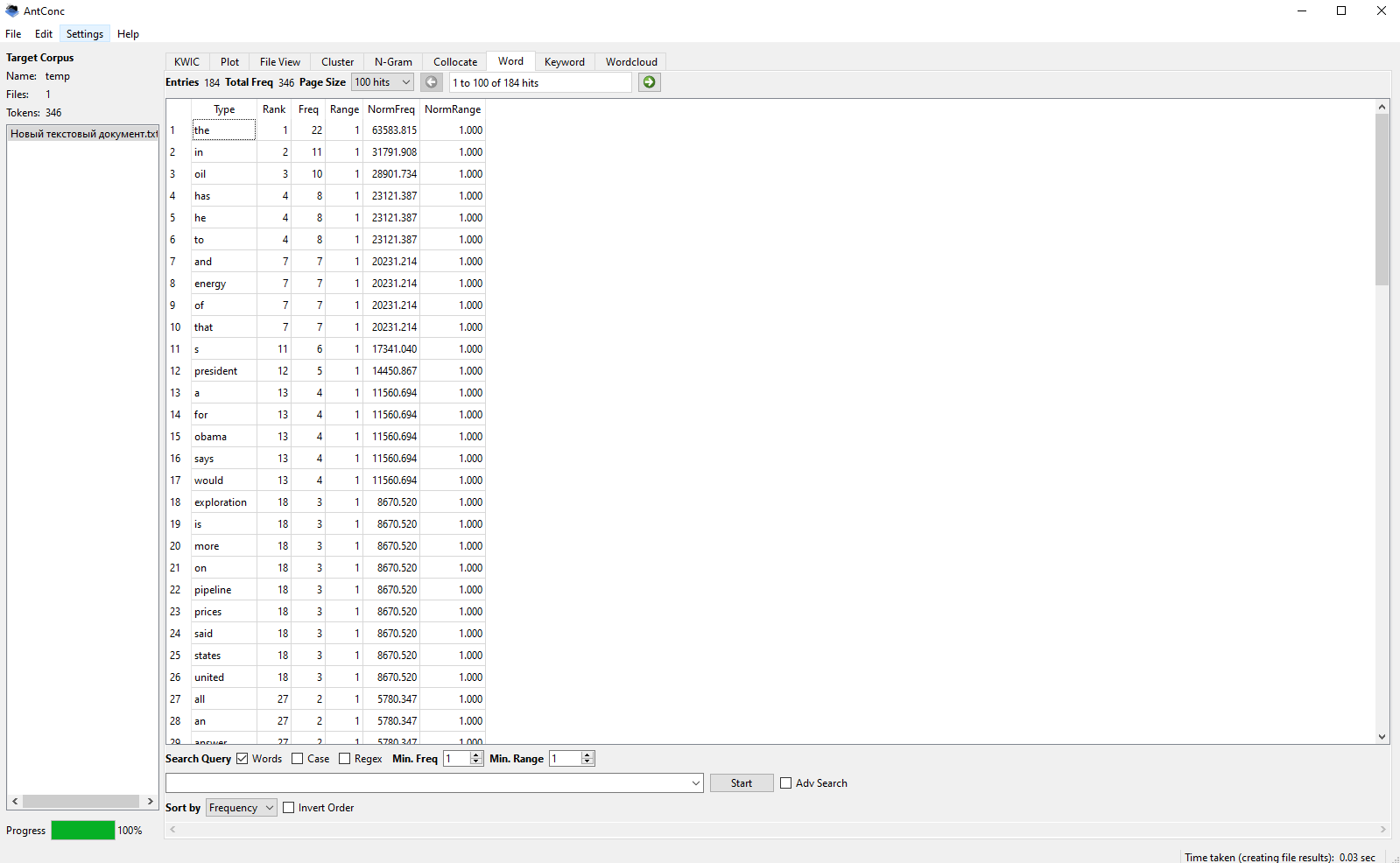
Возьму, как пример статьи на английском языке, следующую статью:

Energy has become an important issue in the American presidential campaign. And gasoline prices, costs that affect almost all Americans, have risen sharply in recent months. For example, former Massachusetts governor Mitt Romney has criticized President Obama’s energy policy. In a newspaper commentary, he said the president has let prices rise while placing too many requirements on the energy industry. One issue in the political debate is a project proposed by the Canadian company Transcanada. Its Keystone XL pipeline would stretch more than three thousand kilometers from Alberta, Canada to several oil processing centers in the United States. If completed, the project would cost about thirteen billion dollars and transport over one million barrels of oil a day. The Republican Party has criticized President Obama for delaying work on the oil pipeline. Republican Doc Hastings serves in the United States Congress. He says the party wants to use all of the nation’s energy resources. Mr. Hastings says that would include use of both the Keystone XL pipeline and oil exploration in coastal waters. He says Republicans in the House of Representatives have passed reforms that break down government barriers to America’s natural resources. And, he says the House has passed legislation that would open coastal areas to new oil exploration and production. In early March, President Obama visited a Daimler truck factory in North Carolina. He told workers that more oil exploration is not an easy answer to the nation’s energy problems. He said the United States has two percent of the world’s oil reserves, but uses twenty percent of the world’s oil. He said this means that, even with increased oil production, there is no easy answer to high gasoline prices. Workers at the North Carolina factory build alternative fuel vehicles. Mr. Obama told them that new sources of energy mean more jobs. The president has called for tax breaks on clean energy vehicles and urged local governments to buy them. He also has asked lawmakers to end four billion dollars in support for the oil and gas industries.

Статья в программе:



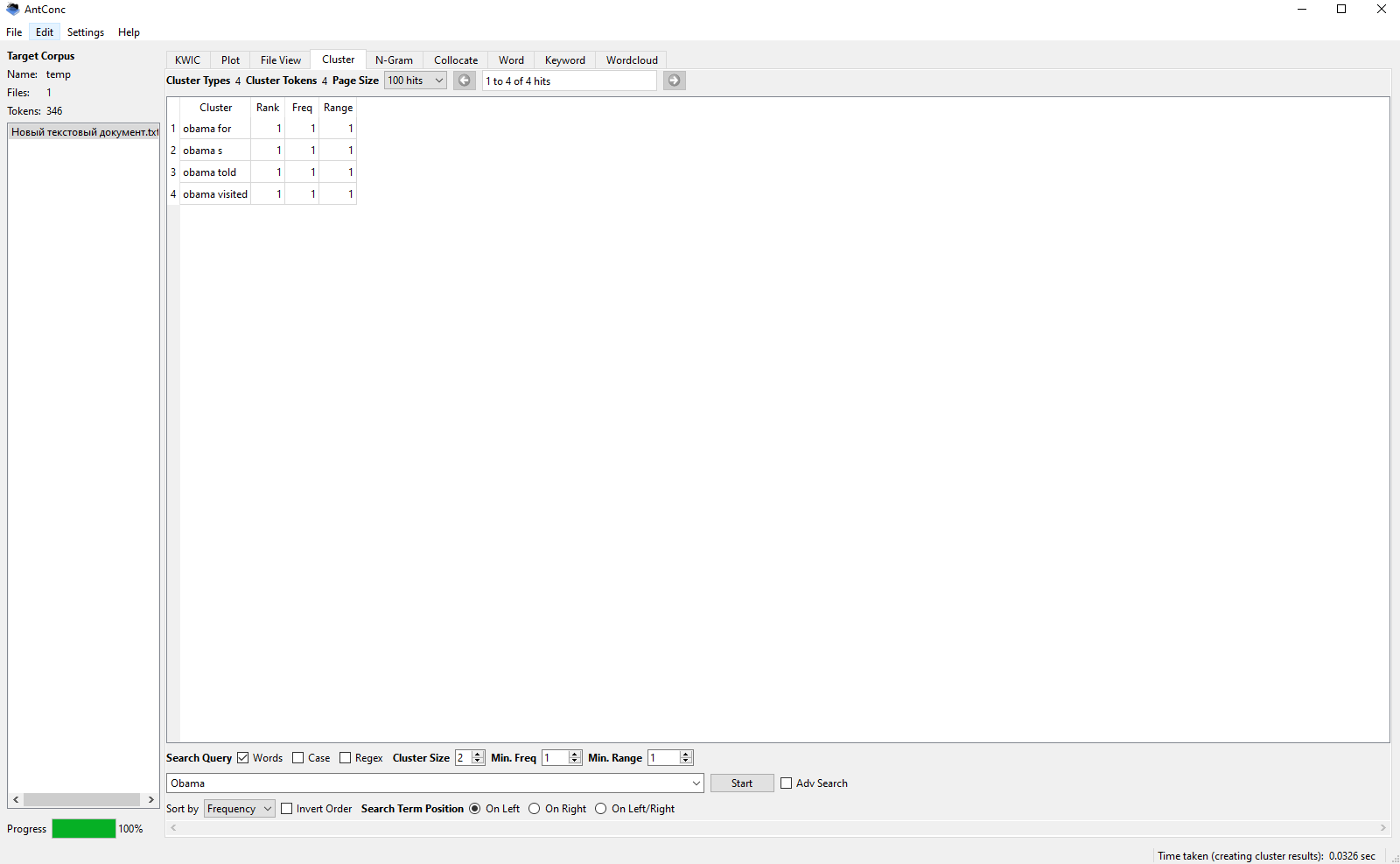
Список частотных лемм:



Список N-грамм:



Список кластеров со словом Obama:



Первые несколько лемм в BNC:







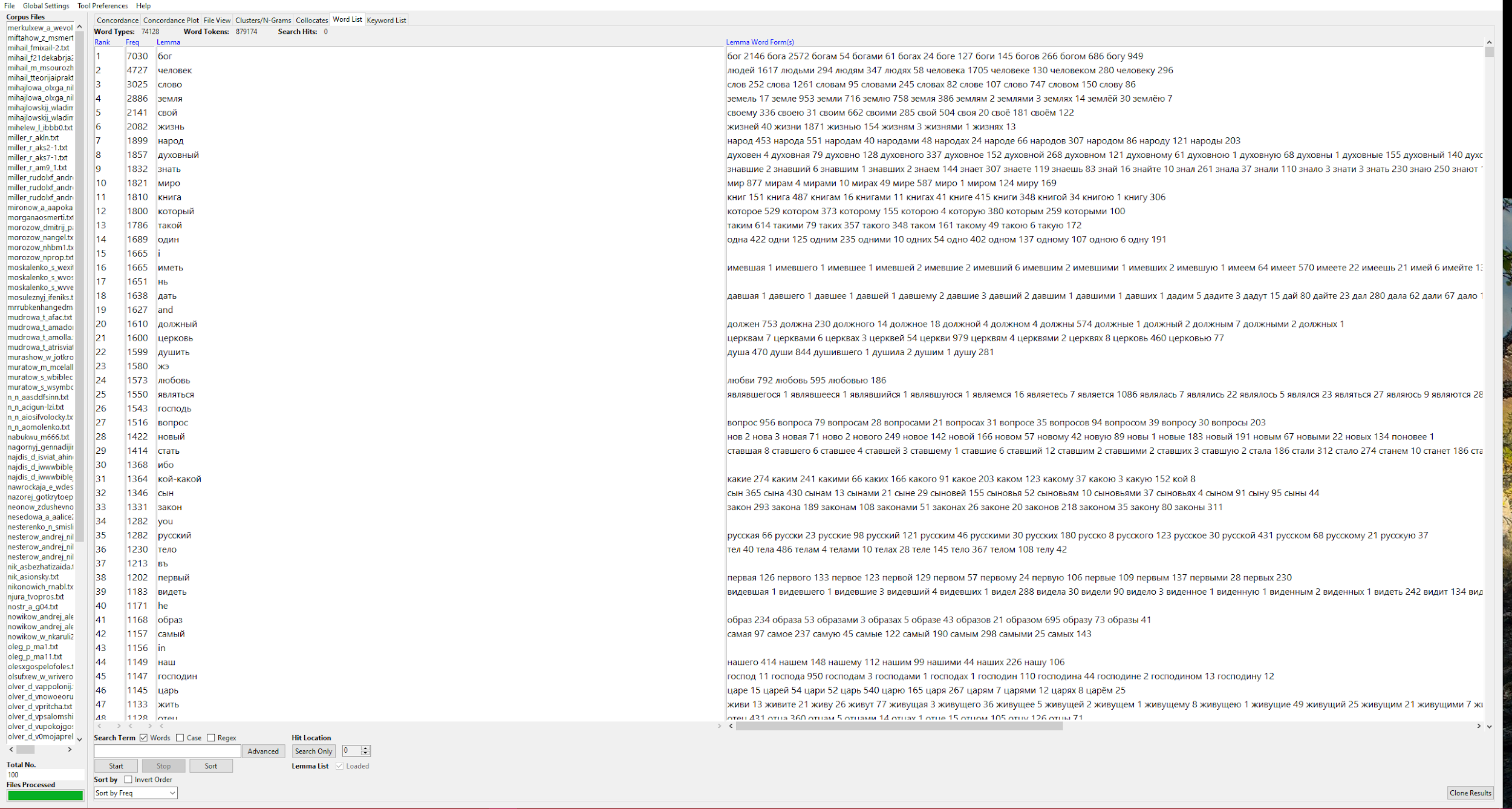




**Уровень 2**

Я выбрал два корпуса текстов на тему мистика и религия. Взял я из архива по 100 текстов. На основе двух тем создал две таблицы в Excel, в каждой из которых был обычный WordList (Столбцы A-C) и с применением списка стоп-слов (Столбцы E-G). Помимо этого, я создал на основе первых 20-ти самых часто встречающихся слов диаграммы, визуализирующие результаты. Было очевидно, что первые позиции без списка стоп-слов занимают непосредственно эти слова, однако, если их отсеять, то можно увидеть, какие самые встречаемые слова в корпусах. В корпусе религиозных текстов на лидирующих позициях были такие слова, как бог, жизнь, человек, церковь, господь, земля, любовь, различные их словоформы и так далее. В то же время в корпусе мистических текстов ситуация вовсе непонятная: самыми популярными словами были кира (непредсказуемое вовсе слово, которое встречается больше остальных), глаза, лили, руки, голову, влад (самое, пожалуй, неожиданное) и так далее. В общем, подводя итог, можно сказать, что если не отфильтровывать стоп-слова, то первые места по частотности в любом тексте будут занимать союзы, артикли, местоимения, предлоги и другие вспомогательные части речи; также я понял, что стандартного списка стоп-слов может зачастую недостаточно, поэтому для каждого текста его лучше дополнять самому. Если же смотреть реальную картину частотности то, конечно, в корпусах на разные темы встречаются разные слова разное количество раз, но стоит уточнить, что с темой религии всё достаточно однозначно, а с мистикой, как не странно, все окутано загадками и непонятно, почему те или иные слова встречаются чаще других.

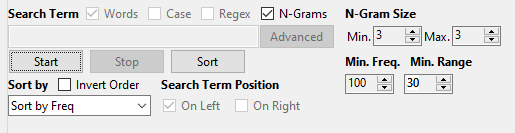
Интерпретированный список лемм:



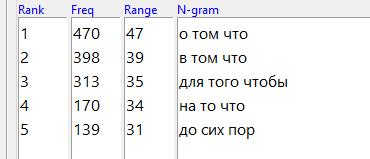
Теперь получим поочерёдно цифру соотношения общего количества слов и числа уникальных слов в каждом корпусе (без удаления стоп-слов, все находятся в равных условиях, так что всё честно, а прогон через список каждого корпуса слишком времязатратен). Для религии она составит , для мистики – , а для профессиональной прозы – . Заметим, что в профессиональной прозе цифра самая наименьшая, а это значит (как я понял), что в прозе тексты менее разнообразны, чем в наших корпусах.

Далее – триграммы.

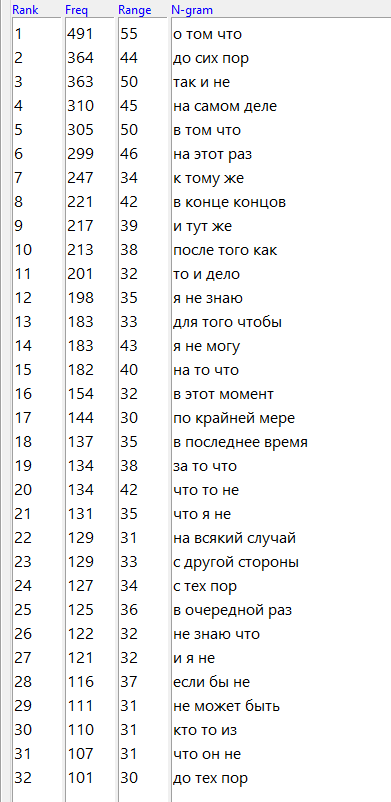
У нас есть общие требования, которые выглядят вот так:



Для текстов про религию имеем следующую картину:



Для текстов же про мистику что-то подобное:



Заметим, что в мистике триграмм больше, но в обоих корпусах это не какие-то отличающие корпус выражения, а достаточно стандартные фразы. Однозначно самой частотной фразой будет “о том что”.

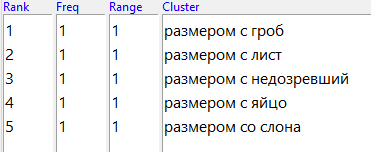
Небольшая часть таблицы самых встречаемых существительных после фразы “размером с”:



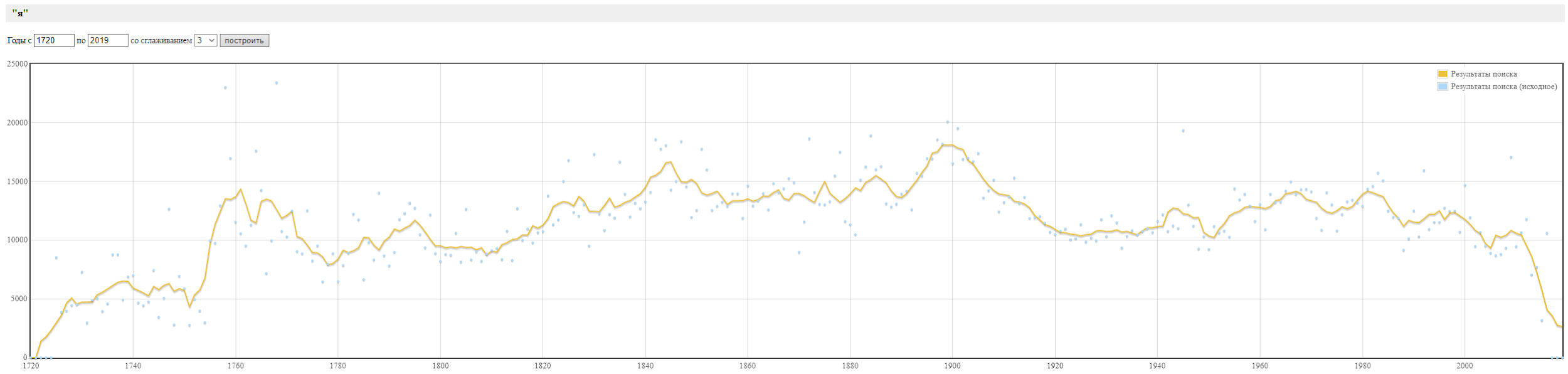
Сравним их с подобными триграммами в наших корпусах. Ситуация с мистикой:



Ситуация с религией:



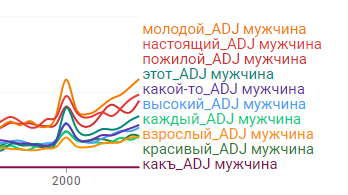
Самые частые триграммы, начинающиеся “размером с” часто совпадали с теми, которые я получил в НКРЯ, например, размером с кулак или с ладонь. Но были и триграммы, характерные для определённого корпуса также, такие как, размером с гроб.

Далее проверим гипотезу, что любительская литература может отличаться более частым употреблением личных местоимений. Возьмём, например, местоимение “я” и найдём, сколько раз оно встречается в наших корпусах. Исходя из таблиц, составленных раннее, можно увидеть, что в текстах про мистику местоимение “я” встречается 33033 раза, а в религии 14602 раза. Следовательно частота появления в корпусе равна 0,014272 и 0,010283 соответственно. Посчитаем также число вхождений слова “я” в поэтическом корпусе: . Также приведу таблицу частотности данного слова по годам, из которой явно видно, что в последние годы в художественных произведениях подобно 1720 году слово я начинает всё меньше встречаться: 

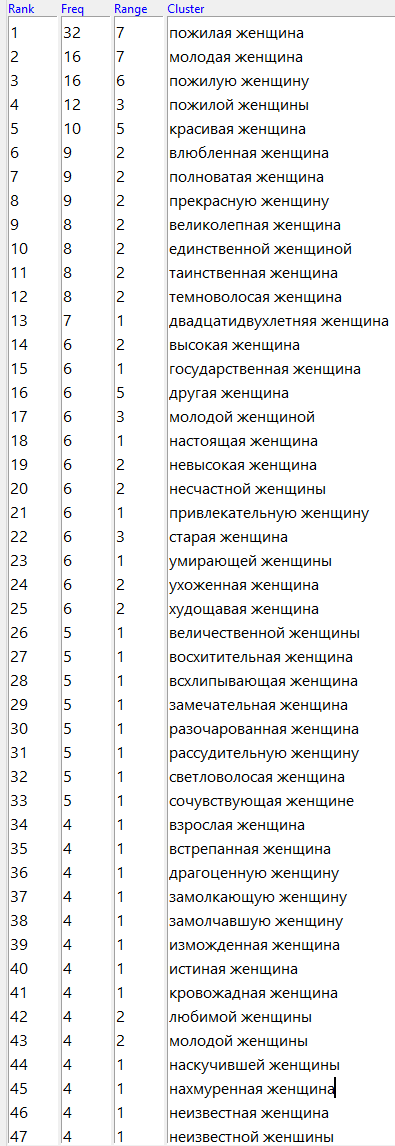
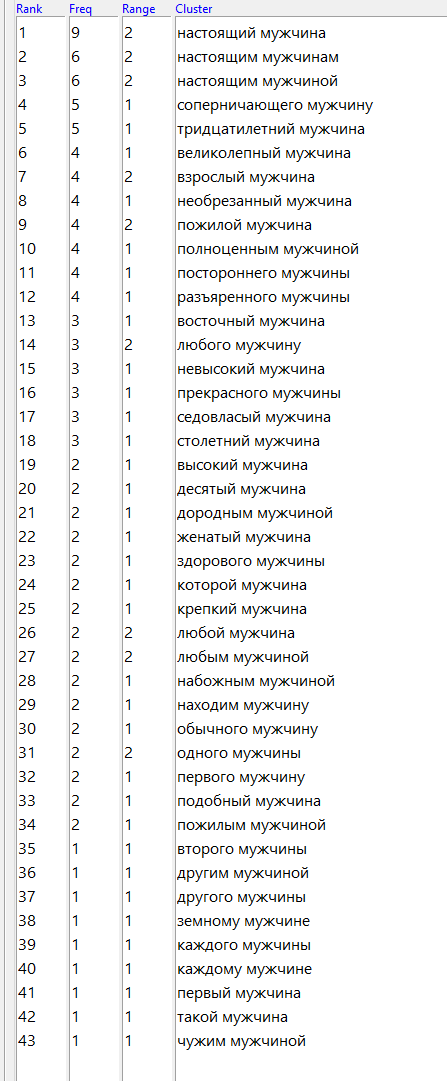
Получив три значения частотности, можно сделать вывод о том, что слово я встречается везде примерно одинаково, где-то больше, где-то меньше, чем в НКРЯ, но в общем и целом, частотность схожая, но тенденция идёт к уменьшению использования данного местоимения.

Далее через НКРЯ узнаем, какие прилагательные чаще всего предшествуют существительному мужчина и женщина, результаты в таблице. Увидим, что общего много, но есть и различия, которые по большей части являются просто словами, свойственными определённому полу: беременная женщина или бородатый мужчина.

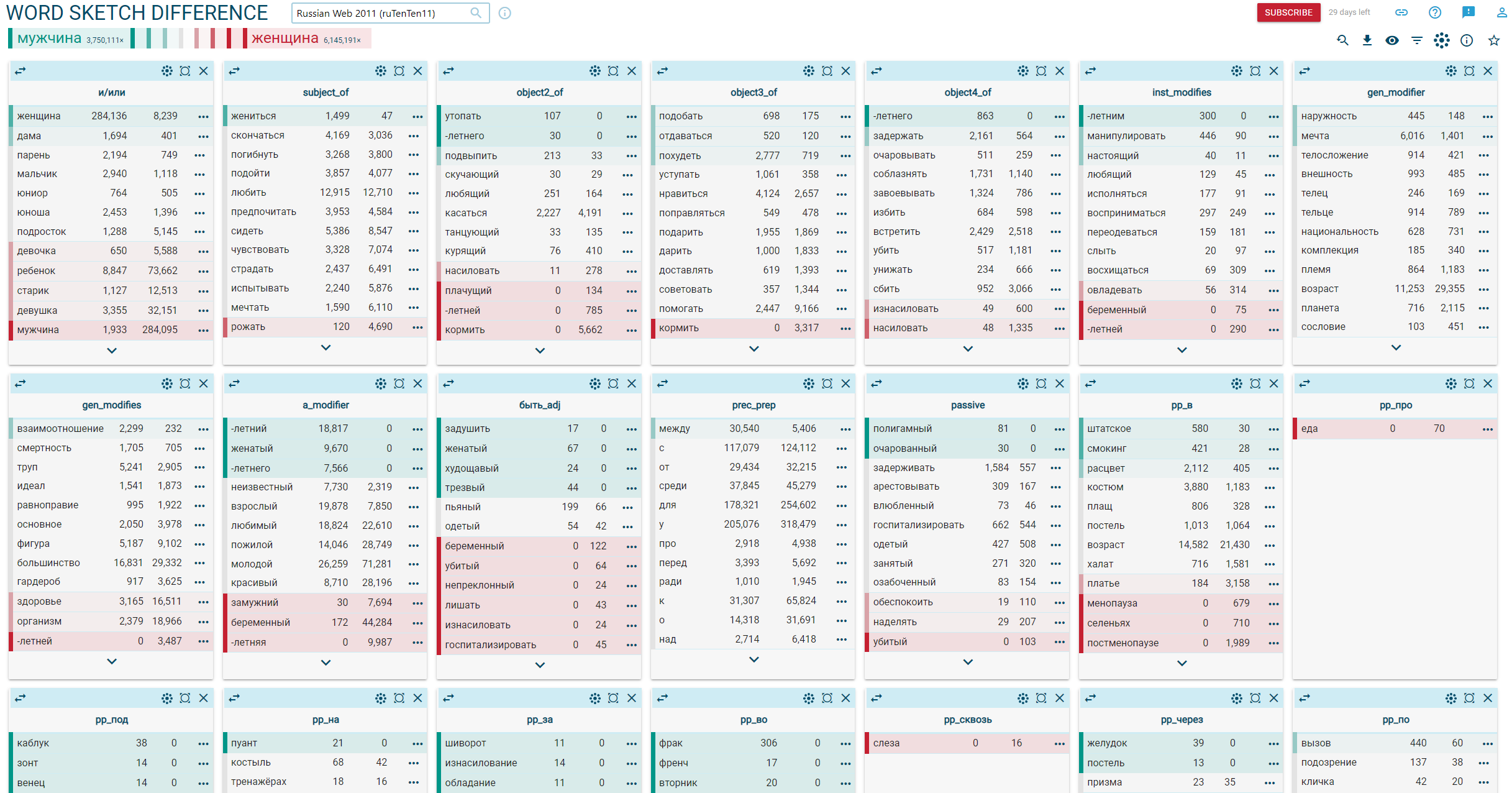
Ситуация с Google Ngram Viewer:



Увидим, что здесь уже присутствуют такие сочетания, как “этот мужчина” или “эта женщина”; “какой-то мужчина” и так далее, что не очень отражает всю картину самых частотных прилагательных перед определённым словом. Вот так дело обстоит в моём религиозном корпусе на примере поиска мужчин (слева) и в моём мистическом корпусе на примере женщин (справа):

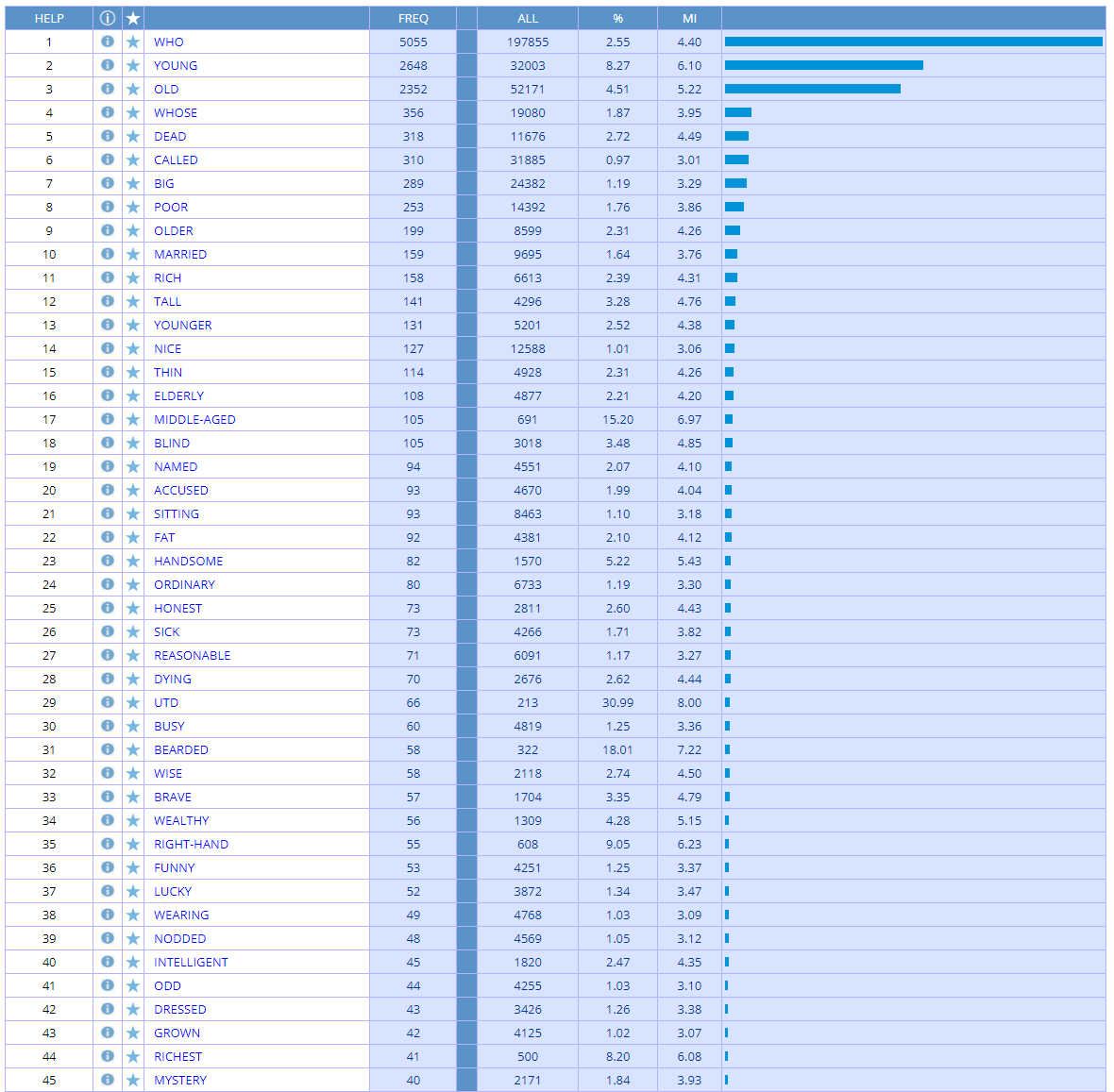


Коллокации слов мужчина и женщина в Sketch Engine:

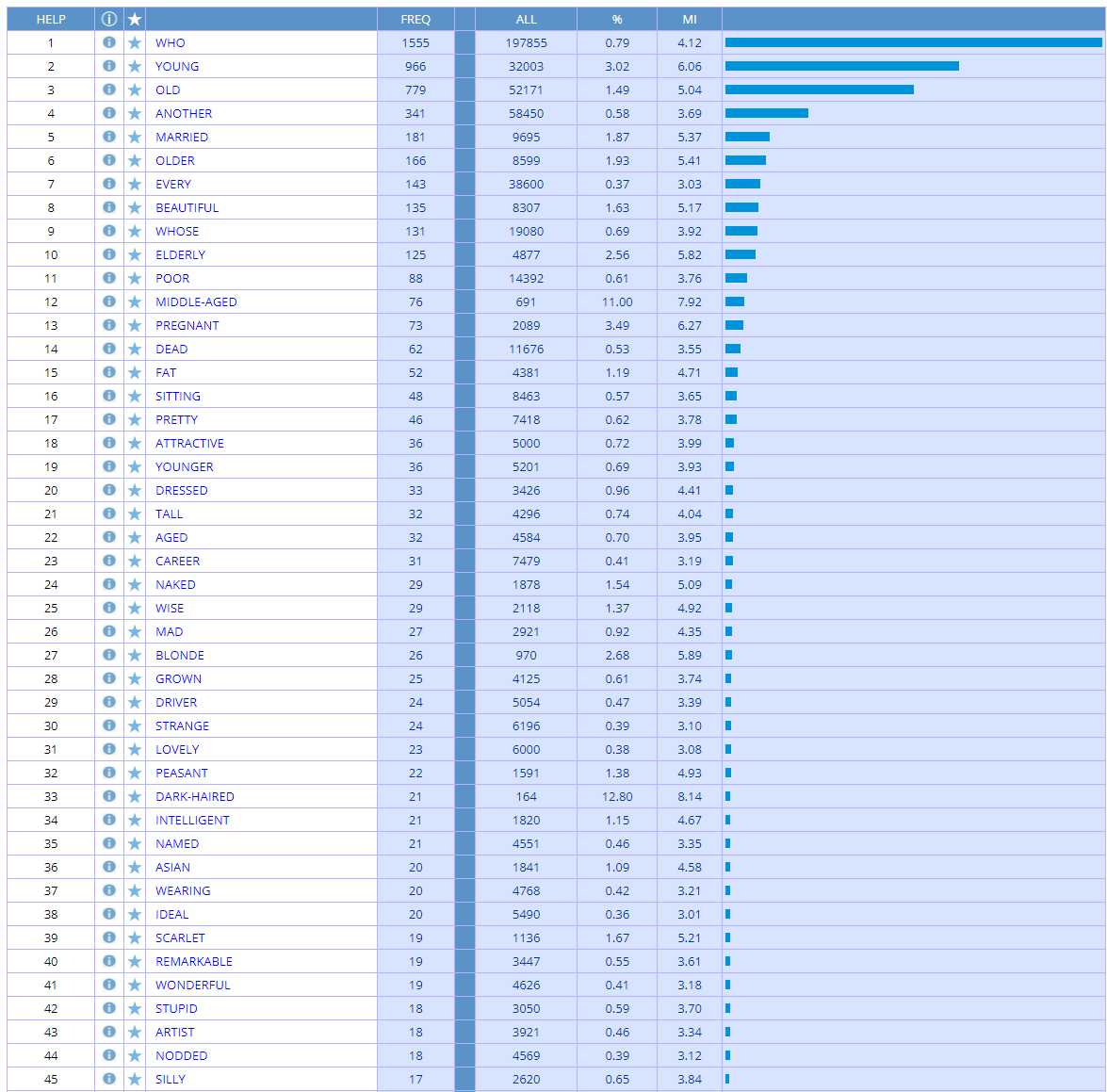


Большая часть различий по прежнему заключается в тех или иных словах, характерных для определённого пола.

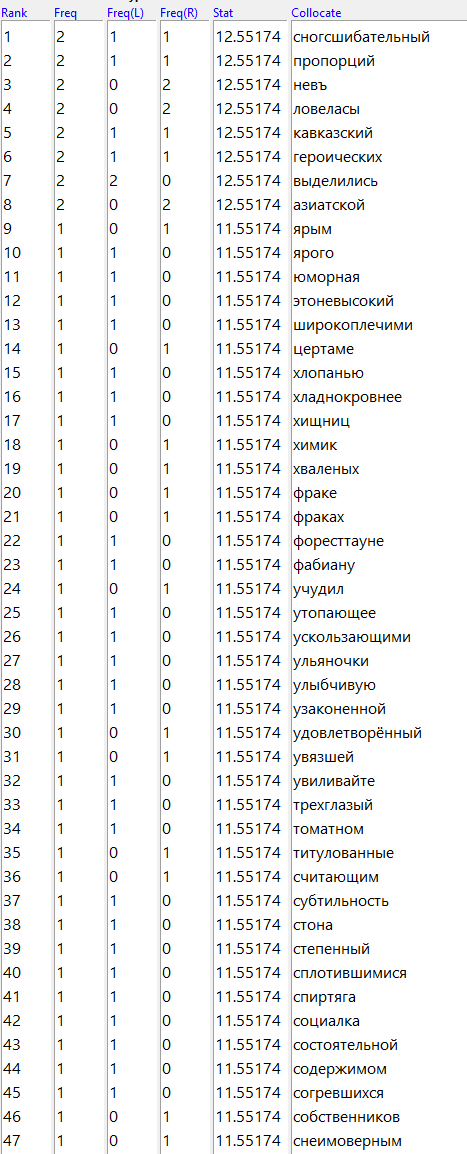
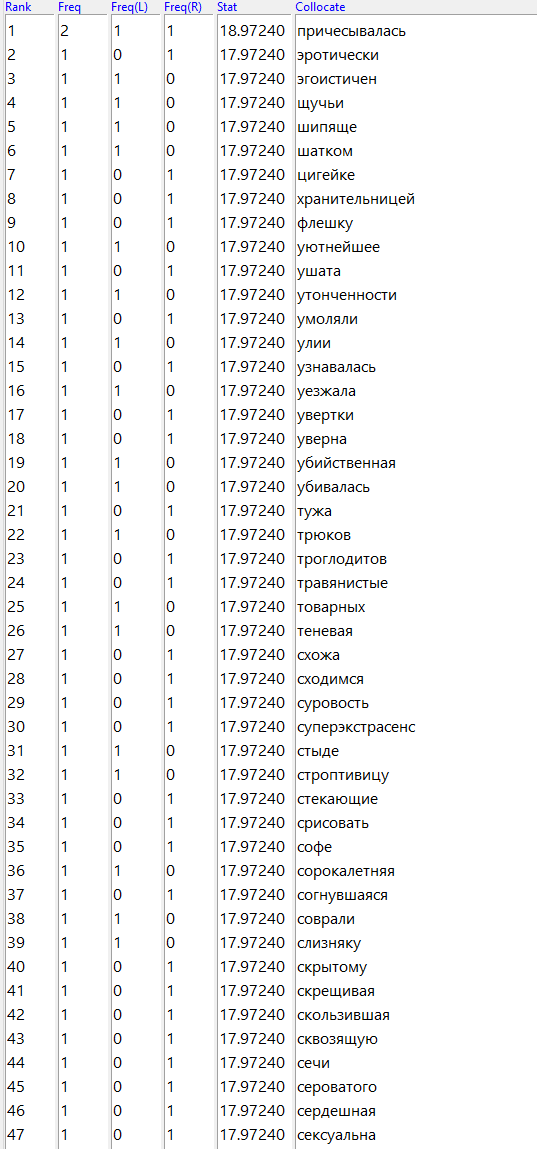
Далее коллокации слова man в BNC:



И соответственно woman:



Пара примеров с ситуацией в моих корпусах:

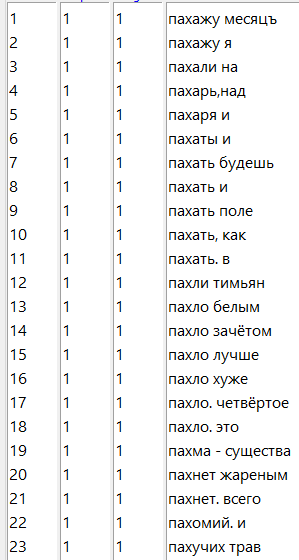
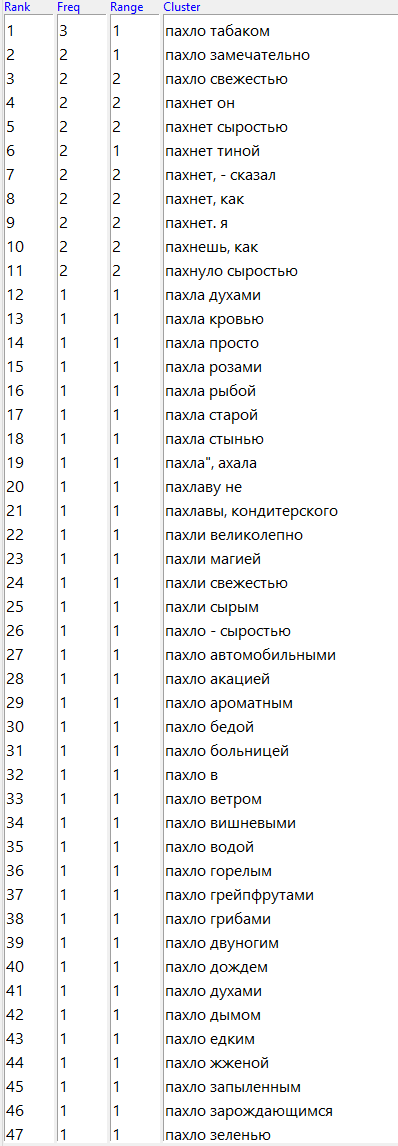


Слова в моих корпусах, которые встречаются со словами мужчина и женщина из-за малого числа текстов редко совпадают с коллокациями в Sketch Engine или BNC, поэтому списки так отличаются, но причиной также может служить и специфика корпусов.

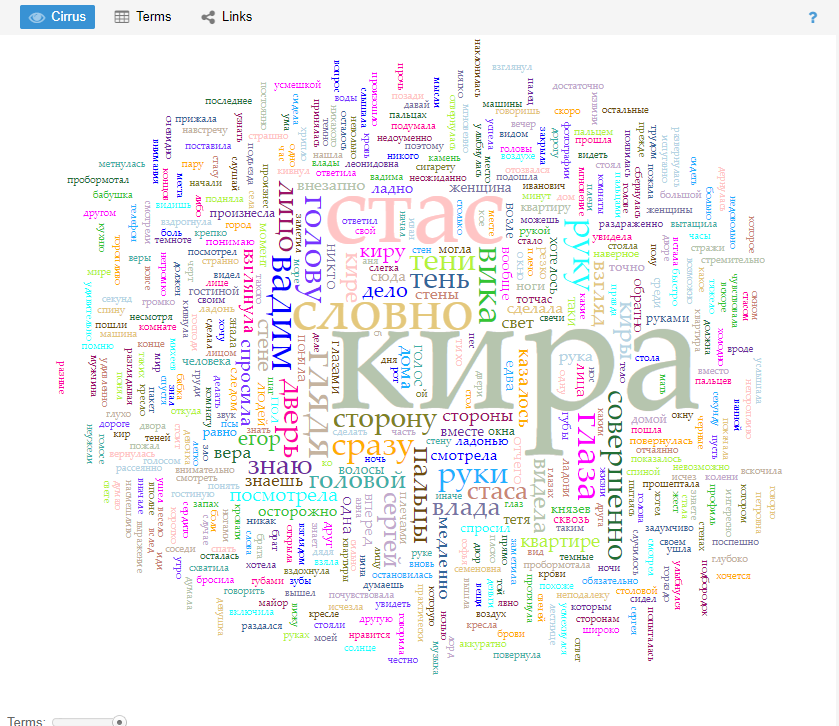
Ниже представлены биграммы с различными словоформами “пахнуть”:



И в моих корпусах (слева мистика, справа религия):



Во втором тексте заметно чаще встречаются слова, которые по значению никак не похожи на пахнуть, да и самих слов сильно меньше. Сравнивать трудно, так как в НКРЯ большая часть коллокаций с наречиями или предлогами, но заметных совпадений не особо много.

И последнее задание, а именно – облако тегов, созданное через сайт Voyant Tools. Для мистики: 

И для религии (взял не самый большой текст, так как самый большой имел недопустимые символы, не кириллица, а не пойми что):

