УДК 811.133.1

НЕКОТОРЫЕ ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ НАУЧНО-ИССЛЕДОВАТЕЛЬСКОГО КОМПЛЕКСА Google Books Ngram Viewer

В.И. Богодист
Ульяновский госуд рственный пед гогический университет им. И.Н. Ульянов , г. Ульяновск

1. Кр тк я х р ктеристик комплекс

Данный сайт – детище фирмы *Google* существует с 16 декабря 2010 г. [4]. Он позволяет проследить встречаемость словоупотреблений и фраз в книгах, опубликованных с 1600 г. по настоящее время (в Google Books сейчас оцифровано 15 млн. книг, это составляет около 10 % всего опубликованного книжного фонда на английском, французском, испанском, русском, немецком, итальянском, китайском языках, а также на иврите. Заметим, что скриншоты (графики), отражающие статистические истории словоформ, слов, словосочеаний, фраз, показывают соответствующую информацию, начиная с 1800 по 2000 год.

В 2012 г. в России также появился аналогичный сервис, позволяющий осуществлять диахронические исследования [5]. Они ведутся в Национальном корпусе русского языка (НКРЯ). Сервис работает на текстах НКРЯ и называется «Графики». Им могут воспользоваться как специалисты в области русского языка и культуры, так и студенты филфака.

2. К к пользов ться сервисом Ngram Viewer?

Обратимся за информацией к некоторым статьям, посвященным этому гигантскому сервису и опубликованному в Интернете. Начнем с рекомендаций В. Сидорова [3] :

– зайти на сайт Google Ngram Viewer;
– в текстовое поле Graph these case-sensitive comma-separated phrases ввести искомое слово (фразу);
– в текстовое поле between ввести год начала цикла отслеживания фразы, а в текстовое поле and – год конца цикла;
– в выпадающем списке from the corpus выбрать язык;
– нажать кнопку Search lots of books;
– система построит график частоты использования слова (фразы);
– под графиком в разделе Search in Google Books будут приведены ссылки для поиска в Google Books.

Этой информации достаточно для получения пользователем первичных данных. Однако при углубленной работе над словарным составом и усложнении целей и задач исследования необходимо

владеть и другими знаниями, которые можно найти в публикациях Захарова В.П. и Масевича А.Ц. [2].

Приведем подробные выдержки из этой статьи, полезные начинающему пользователю сайта, близко к тексту.

1. В публикации рассматриваются возможности системы Google Books Ngram Viewer, осуществляющей поиск и построение графиков встречаемости Ngram в очень больших корпусах текстов на 9 языках. Каждый корпус (кроме итальянского) имеет две версии – 2009 и 2012. Система Google books Ngram Viewer является в настоящее время наиболее мощным инструментом диахронических исследований. Специалистам в области английского языка и студентам отделений английского языка полезно знать, что система также содержит отдельно корпус британского и американского английского языка, корпус всех вариантов английского языка, корпус художественной литературы на английском языке и так называемый гугловский миллион – книги на английском языке с годами издания с 1500 до 2008.

Количественные характеристики корпуса в целом (по данным авторов на 2012 год) следующие: английский язык – 468 491 999 592 словоформы, французский язык – 102 174 681 393 словоформы, испанский язык – 83 967 471 303 словоформы, русский язык – 67 137 666 353 словоформы, немецкий язык – 64 784 628 286 словоформ, итальянский язык – 40 288 810 817 словоформ, китайский язык – 26 859 461 025 словоформ, иврит – 8172 543 728 словоформ (Примечание: указан только общий корпус английского языка).

2. Авторы считают, что система является очень ценным инструментом для диахронических исследований в области лингвистики и истории культуры.

3. Отмечаются лингвистические особенности анализа текстов с помощью данной системы: обрабатываются текстовые единицы (словоформы, а не леммы) и выдаются их статистические характеристики. Последние имеют большое значение для изучения текстовой морфологии. Что же касается приведения словоформ к леммам, то здесь предстоит довольно большой комплекс операций, о которых частично пойдет речь ниже.

4. Приводятся технические рекомендации: как задать программе построение графика омонимичных форм (jaune *adj* и jaune *nom*); как задать программе построение графиков двух-, трех- ,четырех-, пятичленных Ngram. Обратим внимание читателя на возможность использования этой функции для изучения синтаксических сегментов длиной до пяти членов (свободных словосочетаний, фразеологизмов, пословиц и поговорок данной длины).

5. Дополнительная информация: можно ввести одновременно до пяти словоформ (требуется отделение их запятой). На экране будет столько же графиков (скриншотов) разных цветов. При установке флажка в окне case-insensitive система не различает заглавные и строчные буквы, а при его отсутствии различает. Можно указывать временной промежуток исследования (between and – между ... и...) или заказывать текстовые ссылки, фрагменты текстов и даже целые тексты с выделением контрольных. По вертикальной оси графика откладывается относительная частота встречаемости заданной N-граммы в данном году, выраженная в процентах. На горизонтальной оси показаны годы, входящие в заданный временной интервал (см. рис. 1).

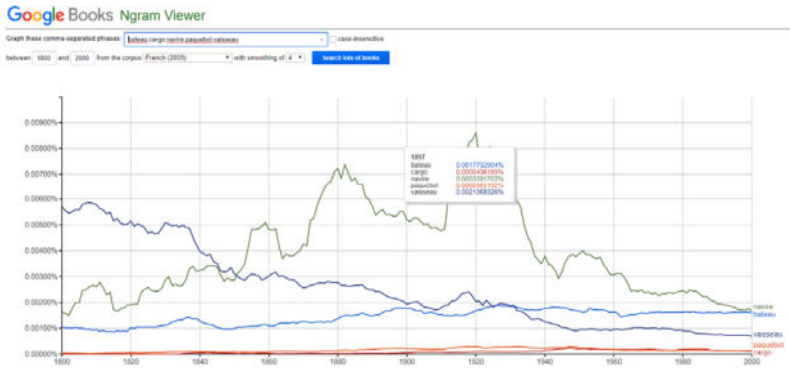

Рис. 1. Скриншот существительных *bateau, cargo, navire, paquebot, vaisseau*

6. Особенности сервиса и представления графиков. Каждая кривая графика маркируется цветом, в конце кривой указывается, какой N-грамме (слову или словосочетанию) она соответствует. Возможно определение координат любой точки графика. Для этого достаточно установить курсор на любую точку над нужным годом. Система в этом случае выдаст сообщение о вертикальной и горизонтальной координатах этой точки для всех кривых. Если же установить курсор непосредственно на кривую, то исследуемая кривая будет выделена.

7. Лингвистические особенности системы: имеется возможность при формулировке условий поиска задавать распознавание заглавных и строчных букв (case sensitive), или игнорировать различие между ними. В системе нет грамматической нормализации лексических единиц, иначе говоря, поиск лексической единицы

(слова или словосочетания) и построение графиков частоты ее встречаемости осуществляется для заданной словоформы.

8. Приводится список тегов для идентификации частей речи. Система предусматривает использование пользовательских тегов для модификации условий построения графиков. Теги этой группы могут применяться изолированно (_NOUN_) в этом случае показывается частота употребления данной части речи, а также могут присоединяться к какому-либо знаменательному слову. Тег _NOUN_, Часть речи Существительное, Действие: программа находит только существительное или субстантивированное прилагательное. Например, «больной» ср. «*Больной н ходится в тяжелом состоянии*» и «*Больной ребенок*» (примеры авторов). В первом случае запрос будет выглядеть следующим образом: больной_NOUN, во втором больной_ADJ. (Все теги частей речи вводятся заглавными буквами без пробелов). Так же работают теги для других частей речи. Полный авторский список тегов включает: Тег _ADJ_ Прилагательное. Тег _VERB_ Глагол. Тег _ADV_ Наречие Тег _PRON_ Часть речи Местоимение Тег _DET_ Часть речи Артикль Тег _ADP_ Часть речи Предлог или послелог Тег _NUM_ Часть речи Числительное Тег_ CONJ_Часть речи Союз Тег _PRT_ Часть речи Частица Тег _INF (Inflections) . Тег «_START_» при необходимости обеспечивает извлечение слова, в том случае, если оно находится в начале предложения. Тег «_END_» позволяет извлечь слово, в том случае, если оно находится в конце предложения. Имеется Тег «_ROOT_=>» используется для поиска глагола, выполняющего роль основного предиката в предложении.

9. Система позволяет строить графики по разным корпусам одновременно. Над кривыми графиков можно производить много операций. Приведем некоторые, которыми может воспользоваться наш начинающий исследователь.

Суммиров ние (сложение) кривых. Операция позволяет суммировать значения каждой точки по оси ординат двух или более кривых. Таким способом можно, например, выяснить вероятность лемм. Для осуществления операции поисковые слова вводятся в окно через знак +, например: *table* + *tables* (*0.00491+0.00117*) .

Вычит ние кривых. Операция позволяет вычитать из значения каждой точки кривой по оси ординат, значение точки другой кривой для той же позиции по оси абсцисс. С помощью этой операции можно представить, насколько частота встречаемости одной N-граммы больше (меньше) другой, и как это различие менялось во времени.

10. Программа дает возможность обрабатывать тексты в дореволюционной орфографии русского языка, так как тексты книг,

изданных до 1919 года (в определенных случаях более поздних изданиях), представлены в старой системе письма, что дает возможность выполнения разнообразных исследований на материале русского языка специалистами, студентами и магистрантами филологического факультета.

11. К сожалению, Google Books Ngram Viewer не может быть помощником при изучении семантики слов, так как она не заложена в его программе. Однако это не умаляет его значения. Сервис может быть полезен при анализе словарного состава языков, изучаемых на факультете иностранных языков и филологическом факультете, так как он позволяет работать с огромными корпусами лексики и решать сложные и трудоемкие задачи. Естественно, во всех случаях исследователь должен проявить находчивость при выполнении действий, прямо не описанных в руководстве по использованию сервиса. Например, программа не идентифицирует омонимы. Для нее *rouge adj* и *rouge nom* – одно и то же слово. То же самое следует сказать о следующих парах слов: *malade adj – malade nom, jaune adj – jaune nom souper v – souper nom и др.* В этом случае нужно действовать следующим образом (см. задание 1).

Задание 1. Определить частоты употребления словоформ, которые приведены в следующих заданиях: malade adj – malade nom jaune adj – Jaune nom souper v – souper nom и др. Использовать теги. Покажем на конкретных примерах ряд процедур. Возьмем пару *malade adj – malade nom*. Для двухтысячного года система выдает для данной словоформы только частоту 0.00333. Нужно установить частоты прилагательного и существительного. Для этого необходимо ввести в окно словоформу следующим образом: malade_ADJ. Получаем цифру 0.00152. Вводим существительное malade_Nom. Получаем отказ – система не нашла запрашиваемую форму. Такой же ответ будет и при вводе формы множественного числа malades. Поэтому, зная общую частоту словоформы и частоту прилагательного, производим операцию вычитания и получаем числовое значение существительного 0,00181.

malade 0.00333
malade adj 0.00152
malade nom 0/00181

Задание 2 *и следующие (ответы для препод в теля). Ан логично 1.*

jaune 0.00146
jaune adj 0.00099
jaune nom 0.00047

Задание 3. *Ан логично 1.*

socialiste 0.00028

socialiste adj 0.00026
socialiste nom 0.00002

***Задание 4**. Ан логично 1.*
rouge 0.00445
rouge adj 0.00248
rouge nom 0.00197

***Задание 5**. Ан логично 1.*
souper 0.0000468
souper v 0.0002358
souper nom 0.0001890

З д ние 6. Для уточнения полноты и ущербности глагольных парадигм, а также употребительности глагольных форм в разных парадигмах можно рассмотреть парадигмы неправильных глаголов, например, глагола ALLER и других в present de l'indicatif, а затем на основании скринщота сделать выводы об их устойчивости в разные временные периоды. *Ср вните эти результ ты. Сдел йте выводы.*

***Тема:** Изучение словосочет ний и фр зеологизмов. Пословицы и поговорки. Это более сложно, но с некоторыми ухищрениями со стороны исследов теля это возможно. Для этого нужно использов ть функцию поиск путем введения в окно до пяти словоформ одновременно.*

***Пример.** Введем изречение Après moi le déluge.* Это крылатое выражение появилось в 1850 г. Его пиковое значение вероятности было *0,0000031277,* в 1866–*0,0000029860,* а далее наблюдается спад до *0,00000008602.*

Это задание можно усложнить, преложив студентам *1) уточнить, кто ск з л эту фр зу и 2) н йти внеконтекстные ст тистические х р теристики к ждой из словоформ, входящих в эту фр зу. 3) можно сост вить и про н лизиров ть в ри нт, приписыв емый м ркизе де Помп дур Après nous le déluge.*

Синонимия. Конкуренция синонимов.

Рекомендуем преподавателю составить похожие задания по стилистике, используя примеры из учебника по французской стилистике К.А. Долинина [1:. 270–272]:

Мелиоративные	***Нейтральные***	***Пежоративные***
–	affirmation	allega
chant	chanson	goualante
–	compagnon	acolyte
–	décret	ukase
économie	–	avarice
serviteur	employé de la mason	valet, laquais

Мелиоративные	*Нейтральные*	*Пежоративные*
–	famille	tribu, couvée
générosité	–	prodigalité
pléiade	groupe	clique
volume	livre	bouquin
–	manoeuvres	agissements,
–	–	machnations
–	paroles, propos	palabres
pieté	–	bigoterie
–	public	galerie
labeur	travail	corvée
–	visage, figure	trogne, mufle,
–	–	gueule
créer	faire	fabriquer
oeuvrer	travailler	trimer
déguster	manger	bafrer
–	parler	palabrer
–	manoeuvrer	mangancer

Задание 7. Следующие синонимы выражают высшую степень положительной оценки [1: 264]. Пользуясь скриншотами, прокомментируйте истории их «жизни»: *Chic, chouette, formid, sensas. Bath, urf, palace, pépère. épatant, extra.*

Задание 8. *Следующие п ры однокоренных слов отлич ются не только суффикс ми, но и зн чениями. Изучите их истории по скриншот м:*

abattage–abattement
adoucissage–adoucissement
ajustage–ajustement
aplatissage–aplatissement
avivage–avivement
barrage–barrement
blanchissage–blanchiement
bon – mauvais
detachage–détachement
expatriation–expatriement
ranimage–ranimation

Выполнение этих заданий не только оживит процесс усвоения стилистики, но и обогатит студентов интеллектуально.

Тема*: Конкуренция суффиксов, н пример,-age и -ment .Т ких п р (троек) довольно много, поэтому з д ний можно придум ть т кже много.*

Тема: *Конкуренция префиксов, н пример, -super и -extra. То же, что и в предыдущей теме.*

Задание 9. *Р ссмотреть следующее множество слов – синонимов*: *bateau , cargo, navire, paquebot , vaisseau .*

Введя эти слова в окно (можно все вместе, но через запятую) получим график (скриншот), на котором пятью цветами изображены пять кривых, фиксирующих положение каждого слова на горизонтальной и вертикальной осях в любое время между 1800 и 2000 годами. Став курсором на любую точку кривой, можно увидеть год и числовое значение относительной частоты его употребления. Следует особо подчеркнуть, что работая со скриншотами, мы имеем дело, во-первых, со словоформами и, во-вторых, с относительными частотами, которыми при необходимости можно воспользоваться для получения абсолютных частот, помня, что объем французского корпуса текстов, использованных для создания веб-сервиса, составляет 102 174 681 393 словоформ.

Скриншот синонимов *bateau, cargo, navire, paquebot, vaisseau* хорошо смотрится в электронном виде, но на бумаге да еще в черно-белом цвете он теряет свою привлекательность. Поэтому вывод следующий: работать нужно с электронным графиком, а важные результаты можно представлять табличным способом.

Пиковые зн чения синонимов в хронологии

Временной период	1801	1850	1901	1950	2000
bateau	0,001272	0,001197	0,002061	0,001713	0,001570
navire	0,002558	0,003277	0,006118	0,003552	0,001925
cargo	0,000005	0,000037	0,000073	0,000216	0,000165
paquebot	0,000048	0,000106	0,000284	0,000246	0,000151
vaisseau	0,000455	0,002132	0,000186	0,000043	0,000087

Задание 10 (к таблице). *Прокомментиров ть т бличные д нные, используя исторические и лингво-культурологические зн ния.*

Тема: Заимствованные слова во французском языке.

Здесь можно выделить подтемы: англицизмы, имеющие эквиваленты (синонимы) во французском языке и англицизмы, не имеющие эквивалентов (синонимов) во французском языке. Аналогичные темы (подтемы) могут выполняться на материале других языков, изучаемых на инфаке.

Задание 11. *Н йти пиковые зн чения вероятностей следующих синонимов, один из которых является з имствов нным словом, и прокомментиров ть*: star f и vedette f., étoile f. По данным PR, первое

слово вошло во французский язык из английского в 1919 году, а сегодня является конкурентом vedette f (1681/конец 19 в.) и étoile f (1549/1849).

Задание 12. *Выявить конкурентноспособность синонимов и прокомментиров ть результ ты.*

coatch–entraineur
fascion–mode
marquetting–marqueterie
standing–permanent
trendy– branché
steck–tranche
soft–doux

Как следует из нашего краткого обзора описываемой системы, она является на сегодняшний день не только непревзойденным по своим информационным возможностям источником, но и облегчает труд исследователя и позволяет анализировать большие словарные массивы, что в принципе невозможно в ручном режиме. Возможности комплекса позволяют выполнять исследования, о которых до последнего времени филологи, культурологи, историки и другие ученые могли только мечтать.

Литер тур

1. Долинин К.А. Стилистика французского языка. – Л.: Просвещение, 1978. – 344 с.
2. Захаров В.П., Масевич А.Ц. Диахронические исследования на основе корпуса русских текстов GoogleBooks Ngram Viewer URL : http://www.allbest.ru/ (Дата обращения ! (15 июля 2018 г.)
3. Сидоров В.П. Как отслеживать частоту применения слов с помощью Ngram Viewer? URL http://netler.ru/articles/ngram-viewer.htm (Дата обращения 15 июля 2018 г.)
4. Google books Ngram Viewer URL http://books.google.com/ngrams (Дата обращения 15 июля 2018 г.)
5. Национальный корпус русского языка URL: http://ruscorpora.ru. (Дата обращения 15 июля 2018 г.)