ДОКУМЕНТАЛЬНЫЕ ИСТОЧНИКИ ИНФОРМАЦИИ

УДК 004.777

В. М. Московкин

Google Books и «культурологические тренды»

На основе аналитико-поискового инструмента Ngram Viewer протестированы термины основных художественных стилей и имена классиков марксизма-ленинизма. Описаны тренды и осцилляции нормализованных частот встречаемости этих терминов и имен, которые нуждаются в дальнейшей интерпретации со стороны культурологов, искусствоведов и историков.

Ключевые слова: *Google Books, Ngram Viewer, culturomics, художественные стили, barocco, gotico, rococo, классики марксизма-ленинизма, Маркс, Энгельс, Ленин, Сталин*

Около семи лет назад компания Google преподнесла грандиозный подарок для исследователей всего мира, запустив сразу три специализированных аналитико-поисковых инструмента: Google Scholar, Google Patents и Google Books. Мы остановимся на последнем.

В декабре 2004 г. компания Google подписала соглашение с пятью крупнейшими библиотеками мира (Нью-йоркской публичной библиотекой, библиотеками университетов Гарварда, Стэнфорда, Мичигана и Оксфорда) по оцифровыванию их книжных коллекций (Google Books Library Project). В настоящее время количество таких библиотек превысило 40. За семилетний период было оцифровано более 5 млн книг (15 млн томов), что составило около 4% всех книг [1]. Темпы, с которыми компания Google начала оцифровывать книги, в свое время вызвали озабоченность со стороны Еврокомиссии и подтолкнули ее к созданию Европейской цифровой библиотеки. Основная масса оцифрованных книг лежит в интервале с 1520 г. по 2008 г. Учитывая, что закон об авторском праве США не распространяет это право на издания, вышедшие в свет до 1923 г., можно получить интервал, в котором находятся книги в статусе общественного достояния: 1520 – 1923 гг.

В работе [2] отмечается, что до 1923 г. в мире было опубликовано около 6 млн книг (18% книг из World Cat). Из нее также узнаем, что более 1 750 000 книг было опубликовано в Западной Европе до 1801 г. Проект Google Books, в первую очередь, нацелен на оцифрование этих книг, т. е. старых и редких книг, на которые истекли авторские права.

Значение этой работы компании Google трудно переоценить, учитывая, что Web покрывает только 20 лет, а печатное слово восходит к Гуттенбергу. Google объявила о грандиозной цели – оцифровать все значимые книги, и, похоже, она этой цели добьется. Многие из этих книг смогут выжить только благодаря их переводу в цифровой вид. Благодаря этой титанической работе будет происходить передача 600-летней человеческой культуры будущим поколениям [1]. И на этом фоне апелляции к каким-то мифическим авторским правам просто блекнут (они или давно истекли или передаются их владельцами компании Google).

Известный научный обозреватель Brian Hayes из «American Scientific» (май-июнь 2011 г.), которого мы уже несколько раз цитировали, пишет, что книги существуют не только для чтения. Имеются другие операции, которые можно делать со словами в книгах. Их можно считать, сортировать, сравнивать, классифицировать, искать закономерности в их распределении и т.д. Все эти способы формализованного анализа позволяют извлекать новое знание из текста так же, как и при его чтении. И это знание гораздо большего масштаба [1]. К этой идее, как пишет Brian Hayes, пришли ученые из Гарвардского университета совместно с аналитиками компании Google. Чтобы распознавать слова в pdf-файле текста книги, они использовали OCR-процесс (оптическое узнавание букв), а учитывая ограничения на авторские права, они из всего текста «нарезали» одно, двух, …, пятисложные словосочетания. В итоге из всего корпуса книг было получено около 500 млрд словосочетаний, которые названы «n-grams» ($1 \le n \le 5$). Для каждого года и каждого «n-gram» определяются перечень и страницы книг, в которых это словосочетание обнаружено, а также его встречаемость. Редкие словосочетания, которые встречаются менее 40 раз в обнаруженных книгах, отсеиваются.

Эта методика с ее приложениями в культурологии и лингвистике была опубликована большим

коллективом авторов из Гарвардского университета и компании Google в январе 2011 г. в журнале «Science» [3]. Развитый здесь, на базе Google Books, аналитико-поисковый инструмент назван Ngram Viewer и был запущен в августе 2010 г. В работе [3] показано, как может использоваться этот инструмент в изучении текстуальных изменений в структуре языка и оценке культурологических трендов. Например, были исследованы сдвиги в балансе между регулярными и нерегулярными глаголами в английском языке.

Авторы работы [3] позиционировали новую область знаний, назвав ее «культуромикой» (*culturomics*), по аналогии с «геномикой» (*genomics*). Точно так же, как крупномасштабные коллекции ДНК-последовательностей приводят к новым биологическим структурам, так и масштабные лингвистические данные (последовательности слов) могут помочь в анализе человеческой культуры и ее трендов. Например, были исследованы изменения в жизненных траекториях знаменитых людей в течение прошедших двух веков. В соответствии с n-gram-анализом было показано, что современные знаменитости становятся известными в более раннем возрасте, и их популярность растет быстрее, но в то же время их и забывают раньше. Другое исследование было посвящено лингвистическим проявлениям цензуры и репрессий на примере Марка Шагала и нацистской эпохи в целом [3]. Brian Hayes проделал вторичные эксперименты с Ngram Viewer и, например, показал, что частота встречаемости немецких слов в английском языке уменьшалась в периоды мировых войн, в то же время частота русских имела пик во время холодной войны.

На конец февраля 2012 г., по нашим оценкам, сделанным с помощью Google Scholar, появилось около 110 статей, в названия которых входит термин Ngram Viewer.

С целью показать возможности использования этого инструмента для отечественных исследователей мы также провели серию экспериментов. Посмотрим вначале, как ведут себя тренды встречаемости в английском языке терминов:

barocco (ит., худ. стиль конца XVI – сер. XVIII вв.), его аналога в английском и французском – *baroque*;

gotico (ит., худ. стиль, сер. XII – XV–XVI вв.), его аналоги в английском, немецком и французском языках – *gothic, gotik, gothique*;

rococo (фр., худ. стиль, 1-я пол. XVIII в.).

Записывая через запятые эти термины в поисковой строке Ngram Viewer, мы сразу же получаем временные графики нормализованных частот встречаемости этих терминов. Такая частота для произвольного словосочетания, состоящего не более чем из пяти слов, в заданный момент времени определяется отношением его встречаемости к встречаемости всех словосочетаний данной размерности, выраженным в процентах. Этот показатель вводится с целью нейтрализации «лингвистической» инфляции [1, 3].

В нашем случае на интервале 1700 – 2008 гг. мы получили три графика для терминов *baroque, gothic* и *rococo* (рис. 1). Графики нормализованных частот встречаемости остальных четырех терминов были близки к нулю и поэтому не показаны. Эти же частоты для терминов *gotico, barocco* и их аналогов на интервале времени 1520 – 1700 гг. также были нулевыми. Видим всплеск нормализованной частоты встречаемости термина *rococo* в период его зарождения в начале XVIII в. (во время регентства Филиппа Орлеанского, 1715 –1723 гг.), потом она становится близкой к нулю на фоне быстрого роста популярности термина *gothic* в период с 1720 г. по 1830 г.

В этот же период близкой к нулю была и нормализованная частота встречаемости термина *baroque*. Таким образом, на фоне быстрорастущего интереса к готике частота встречаемости терминов *baroque* и *rococo* в англоязычной литературе периода с 1720 г. по 1830 г. была близка к нулю. Наоборот, спад интереса к готике (1830 – 1920 гг.) обусловил стабильный рост нормализованной частоты встречаемости терминов *baroque* и *rococo* приблизительно до 1960 г. При этом, если до 1920 г. рост их частот был приблизительно одинаковым, то после этого года произошел резкий рост нормализованной частоты встречаемости термина *baroque*.

В поле франкоязычной литературы нормализованная частота встречаемости термина *baroque* вела себя приблизительно так же: медленный рост до 1920 г. и быстрый рост в период с 1920 г. по 1960 г. Эта же частота для термина *gothique* с 1700 г. росла с 10–15-летними осцилляциями вплоть до 1840 г., далее в течение 100 лет наблюдались приблизительно установившиеся колебания этой частоты, а позднее эти колебания шли с затухающим трендом вплоть до нашего времени (рис. 2).

Более сложная осциллирующая динамика нормализованных частот встречаемости рассматриваемых терминов наблюдалась в поле немецкоязычной литературы. Здесь, начиная с конца 30-х гг. XIX в. наблюдаются непрерывные графики нормализованных частот встречаемости для всех семи терминов, многие из которых осциллируют с возрастающим трендом вплоть до конца 90-х гг. XX в. Во второй половине этого века превалируют нормализованные частоты встречаемости терминов *baroque* и *gothic* (рис. 3).

В поле испаноязычной литературы непрерывные графики рассматриваемых частот возникают с 1840 г. для пяти терминов *barocco, baroque, gothic, gothique* и *rococo*. Здесь во второй половине XX в. так же, как и для этих терминов в англо- и немецкоязычной литературе, значительно доминировала нормализованная частота встречаемости термина *baroque* (рис. 4).

Следует предположить, что специалисты по культурологии и искусствоведению смогут квалифицированно интерпретировать особенности поведения вышеуказанных культурологических трендов и осцилляций.

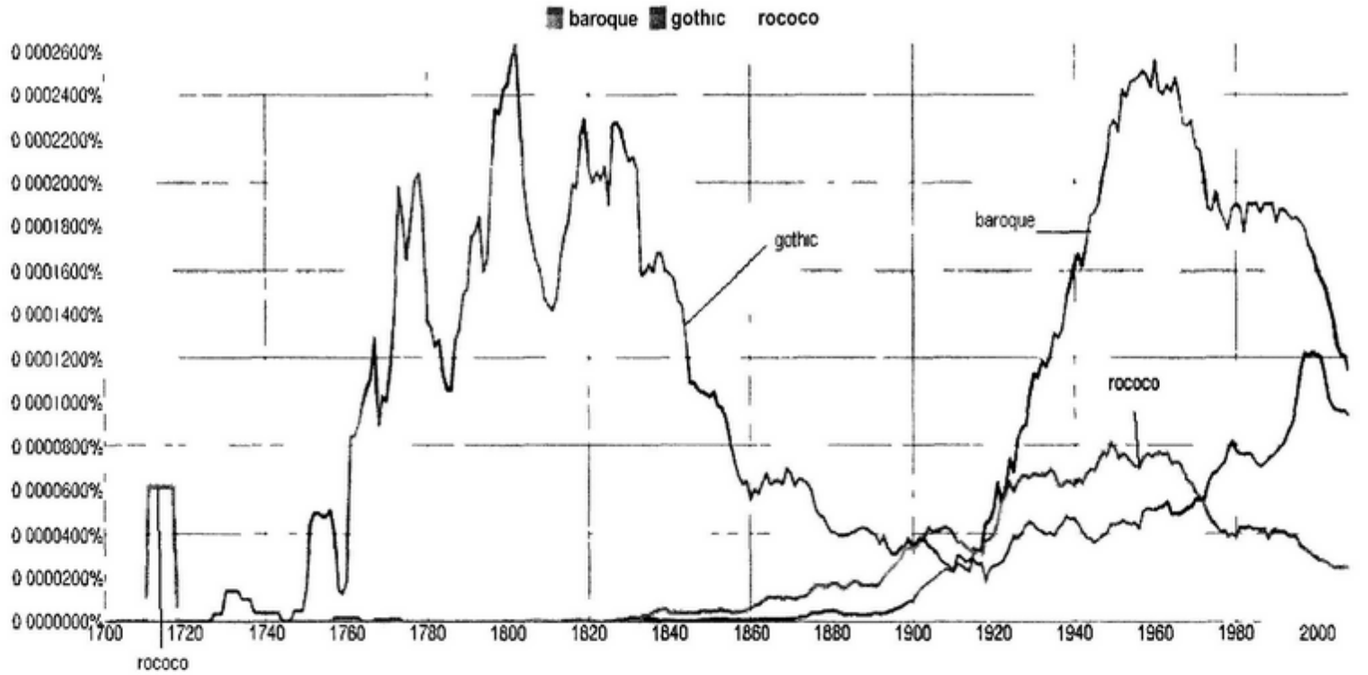

Рис 1 Нормализованные частоты встречаемости терминов художественных стилей в англоязычной литературе

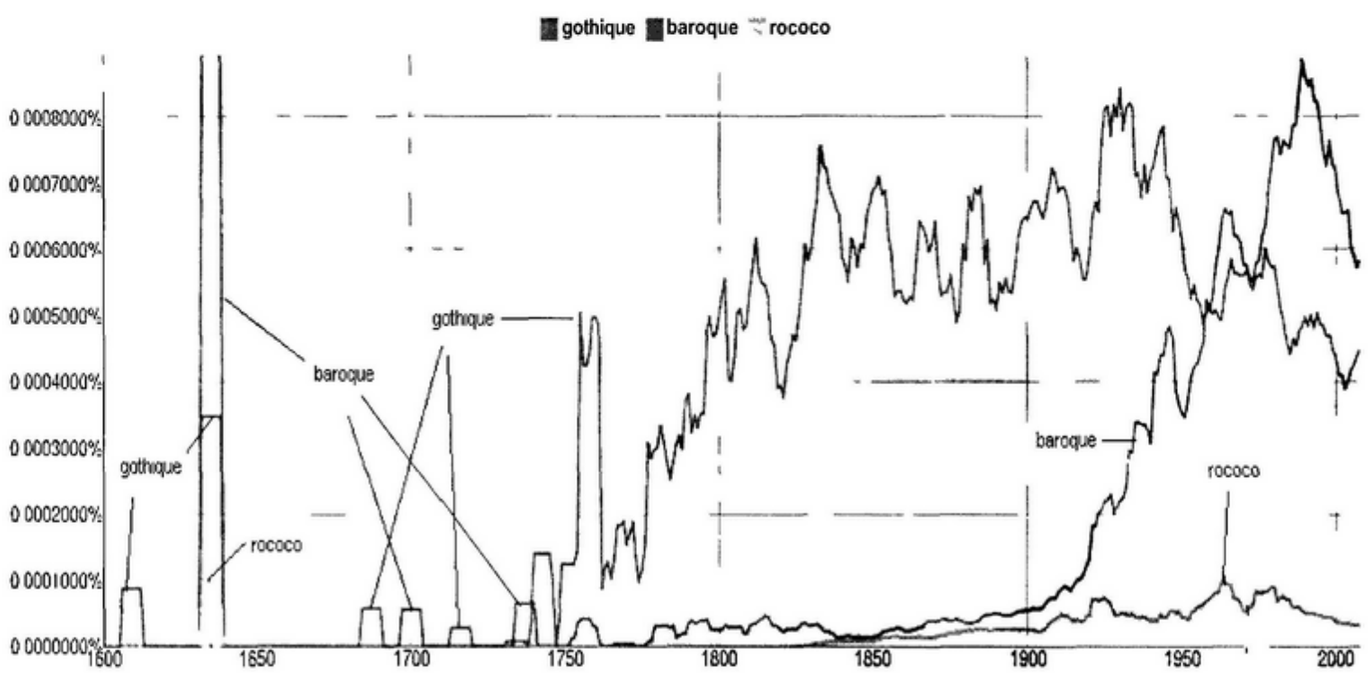

Рис 2 Нормализованные частоты встречаемости терминов художественных стилей в франкоязычной литературе

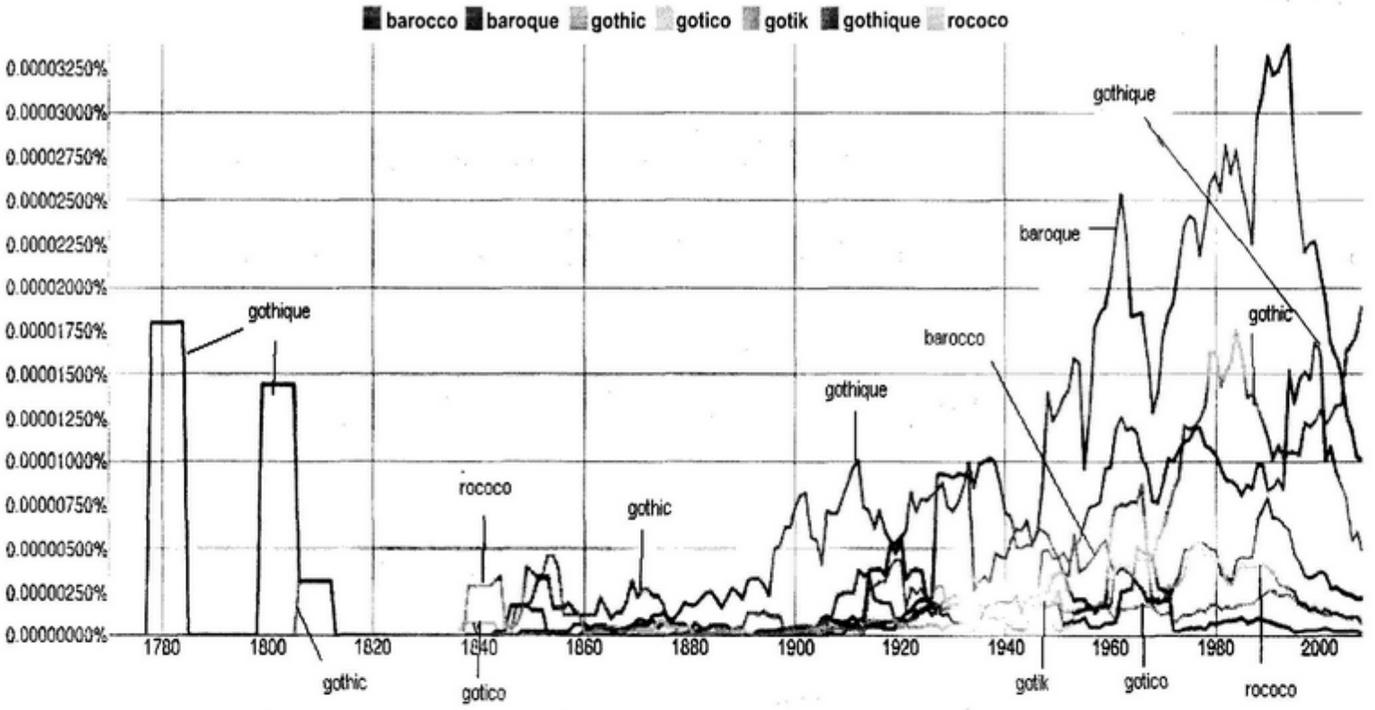

Рис. 3. Нормализованные частоты встречаемости терминов художественных стилей в немецкоязычной литературе

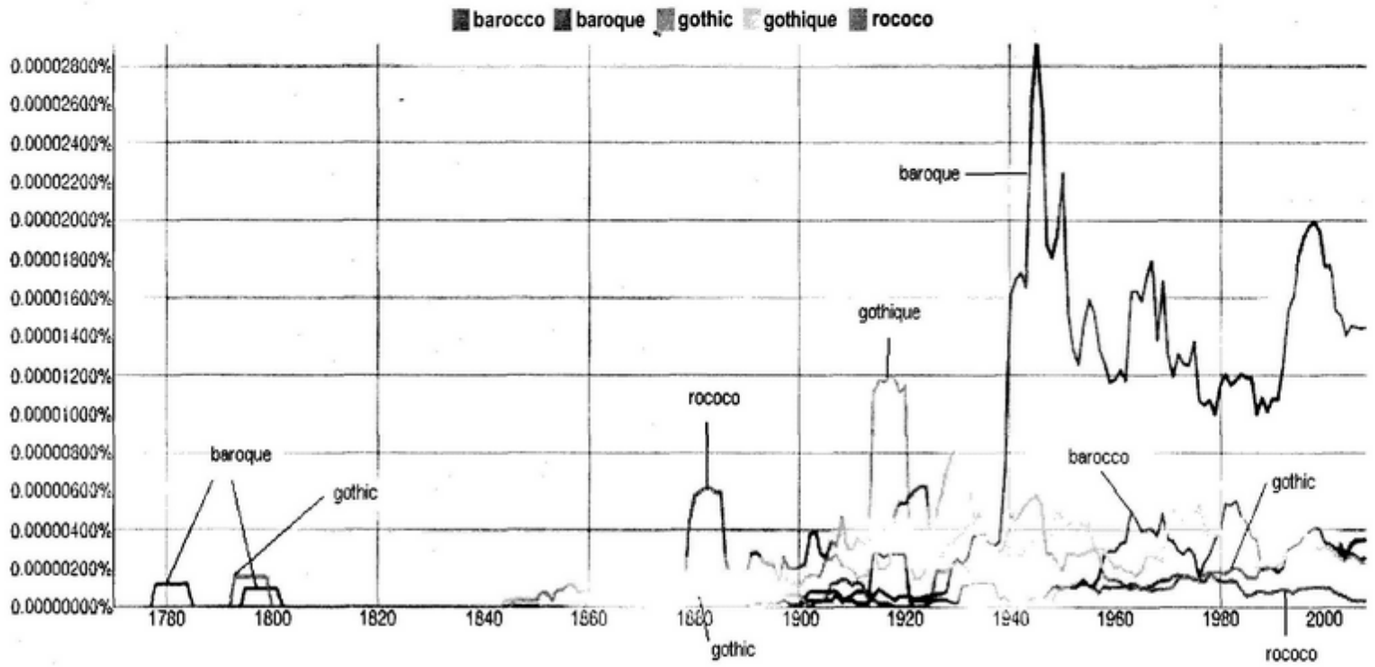

Рис. 4. Нормализованные частоты встречаемости терминов художественных стилей в испаноязычной литературе

В качестве второго эксперимента мы провели тестирование имен классиков марксизма-ленинизма – *Karl Marx, Marx, Friedrich Engels, Engels, Vladimir Lenin, Lenin, Joseph Stalin, Stalin* – на пяти языках: английском, немецком, французском, испанском и русском (рис. 5 – 9).

В случае английского языка соответствующие графики нормализованных частот встречаемости приведены на рис. 5, где видно, что популярность имени *Маркс* начинается с момента издания первого тома «Капитала» (1867 г.), популярность имени *Энгельс* – позже, причем в их пике популярности (1975 – 1980 гг.) популярность имени *Маркс* была приблизительно в три раза выше (0,0025%/0,0008%=3,13), чем популярность имени *Энгельс*. Обращает на себя внимание период с 1936 г. по 1955 г. (вторая мировая война и период восстановления), когда популярность имен *Маркс* и *Энгельс* была стабильной, и последующий за этим период роста их популярности (до 1975 – 1980 гг.), который можно связать с ростом национально-освободительных движений во всем мире и распадом колониальной системы. С конца 70-х годов XX в. наблюдается постепенный спад популярности этих имен (рис. 5).

Динамика изменения популярности имен *Ленин* и *Сталин* намного сложнее. Резкий рост популярности первого начался с 1915 г., второго – с 1923 – 1924 гг. Отметим, что учение «марксизм-ленинизм» было предложено советской партийной верхушкой в 1923 г., оно оставалось действенным, пока был жив Сталин, позднее Хрущев обрушил легитимность этого учения.

Рост популярности имени *Сталин* наблюдался вплоть до смерти его носителя, потом пошел спад с локальными всплесками популярности в 1962 и 1988 гг. Рост популярности имени *Ленин* наблюдался до 1940 г., потом произошел спад до 1948 г., далее рост до 1962 г. и необратимый спад вплоть до наших дней (рис. 5).

Отметим, что особенности поведения нормализованных частот встречаемости имен *Marx* и *Engels* остаются приблизительно одинаковыми для всех четырех иностранных языков (рис. 5 – 8). Популярность имен *Маркс* и *Энгельс* на их пике возрастает последовательно при следующих переходах от языка к языку: английский→испанский→французский→немецкий. Так, на пике популярности имени *Marx* (англоязычная литература – 1980 г., немецкоязычная – 1975 г.) его популярность в немецкоязычной литературе была в 6,4 раза больше по сравнению с англоязычной (0,016%/0,0025%=6,4). Интересная особенность поведения нормализованных частот встречаемости имен *Marx, Engels* и *Lenin* наблюдалась в немецкоязычной литературе. Все они имели четкий локальный минимум в 1940 г. (рис. 7).

При тестировании русскоязычных имен классиков марксизма-ленинизма, мы наблюдаем совершенно другую ситуацию (рис. 9). Популярность имени *Ленин* преобладала вплоть до распада СССР. Это была критическая точка, в которой растущая популярность имени *Сталин* сравнялась с падающей популярностью имени *Ленин*. Наблюдались две четкие волны популярности имени *Сталин*: 1920 – 1959 гг.; после 1985 г. При общем доминировании популярности имени *Ленин* в русскоязычной литературе с 1915 г. по 1992 г., наблюдались ярко выраженные пики его популярности в 1939 – 1942 гг. (0,017 %), 1968 – 1970 гг. (0,024 %). На всем рассматриваемом промежутке времени (1915 – 2011 гг.) средняя нормализованная частота встречаемости имени *Маркс* (0,004 %) была приблизительно в два раза выше, чем имени *Энгельс*. На рис. 9 дополнительно приведен график нормализованной частоты встречаемости имени *Гитлер*. Всплеск его популярности наблюдался в годы 2-й Мировой войны. Отмечается слабый рост популярности этого имени после распада СССР вместе с ростом популярности имени *Сталин*.

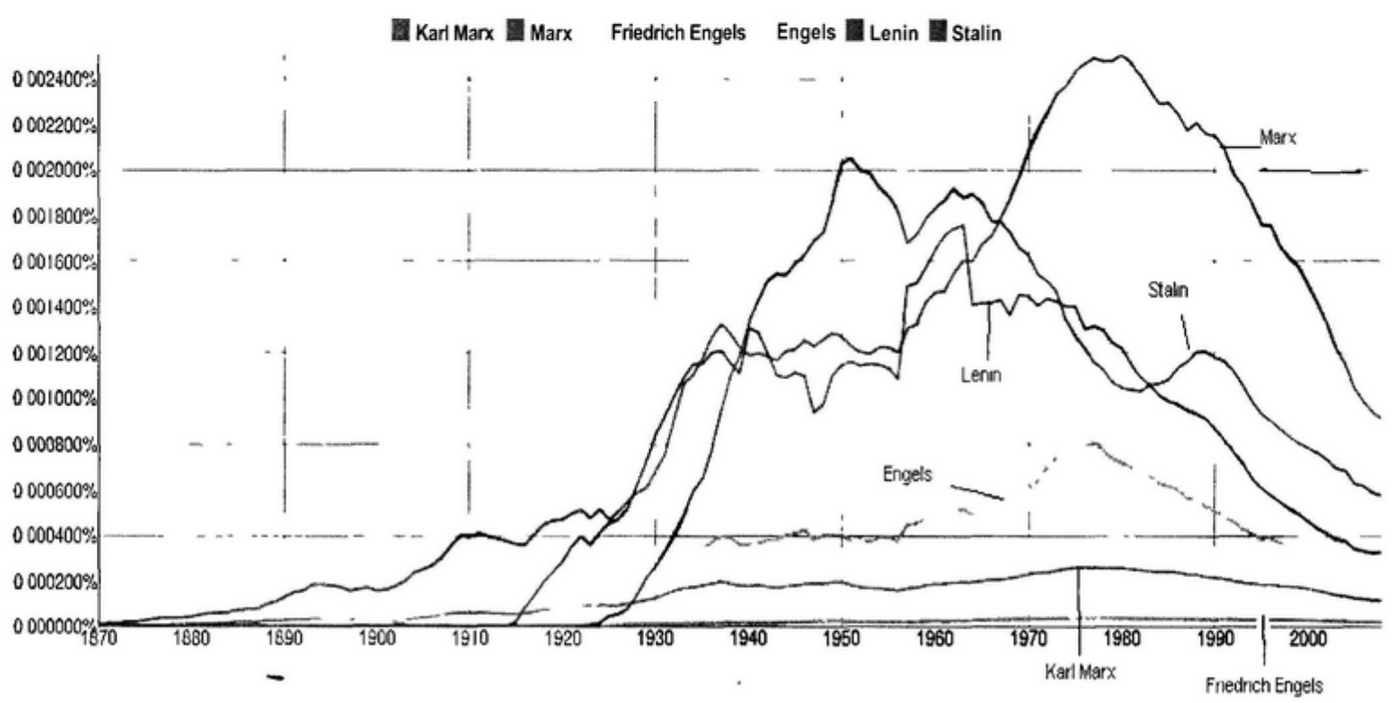

Рис. 5. Нормализованные частоты встречаемости имен классиков марксизма-ленинизма в англоязычной литературе

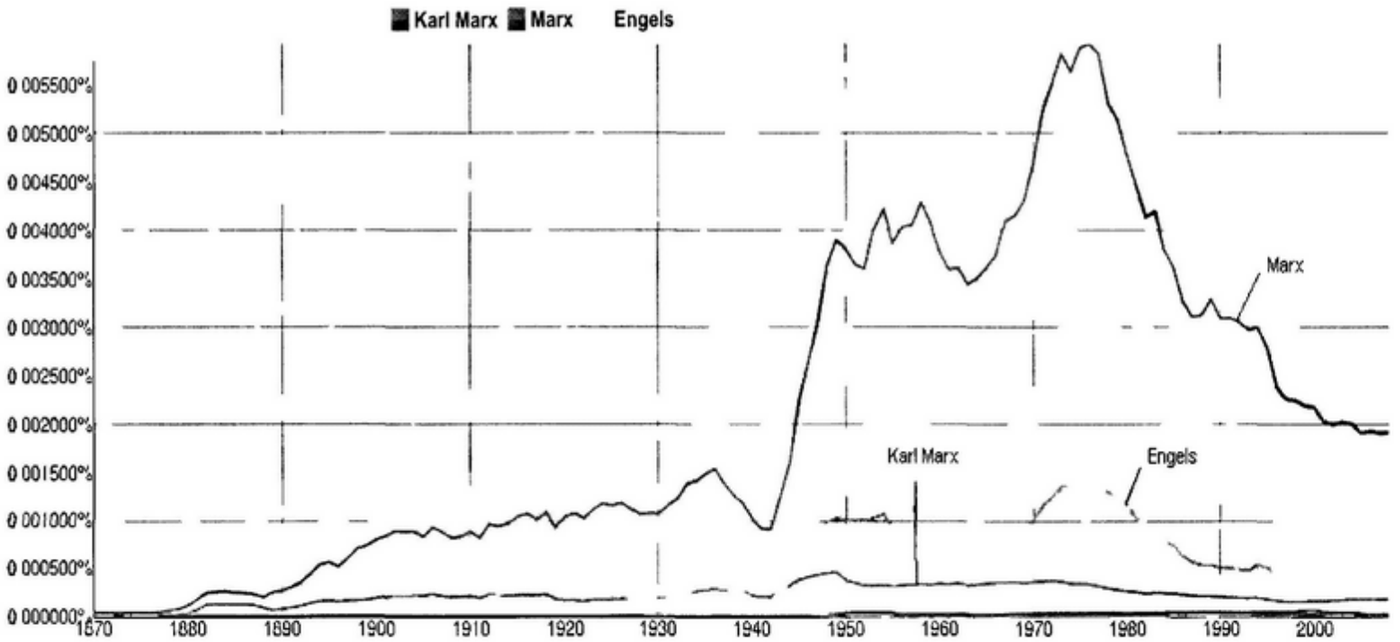

Рис 6 Нормализованные частоты встречаемости имен классиков марксизма-ленинизма в франкоязычной литературе

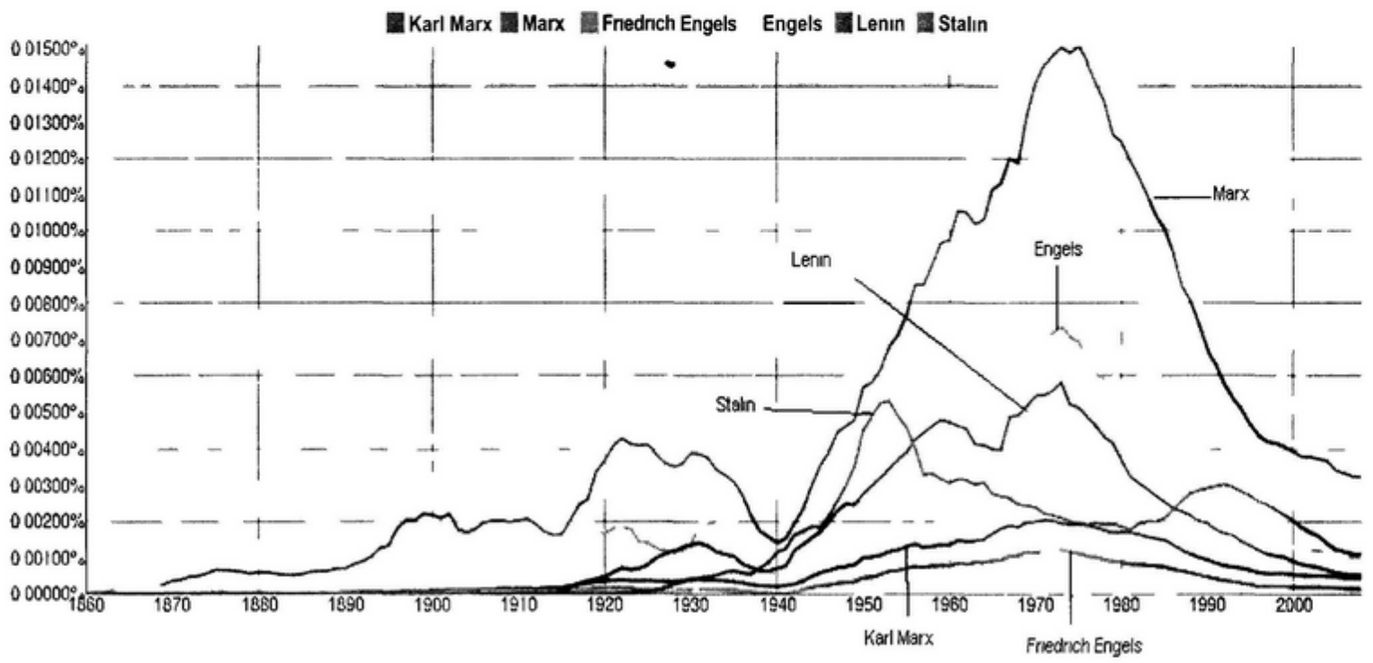

Рис 7 Нормализованные частоты встречаемости имен классиков марксизма-ленинизма в немецкоязычной литературе

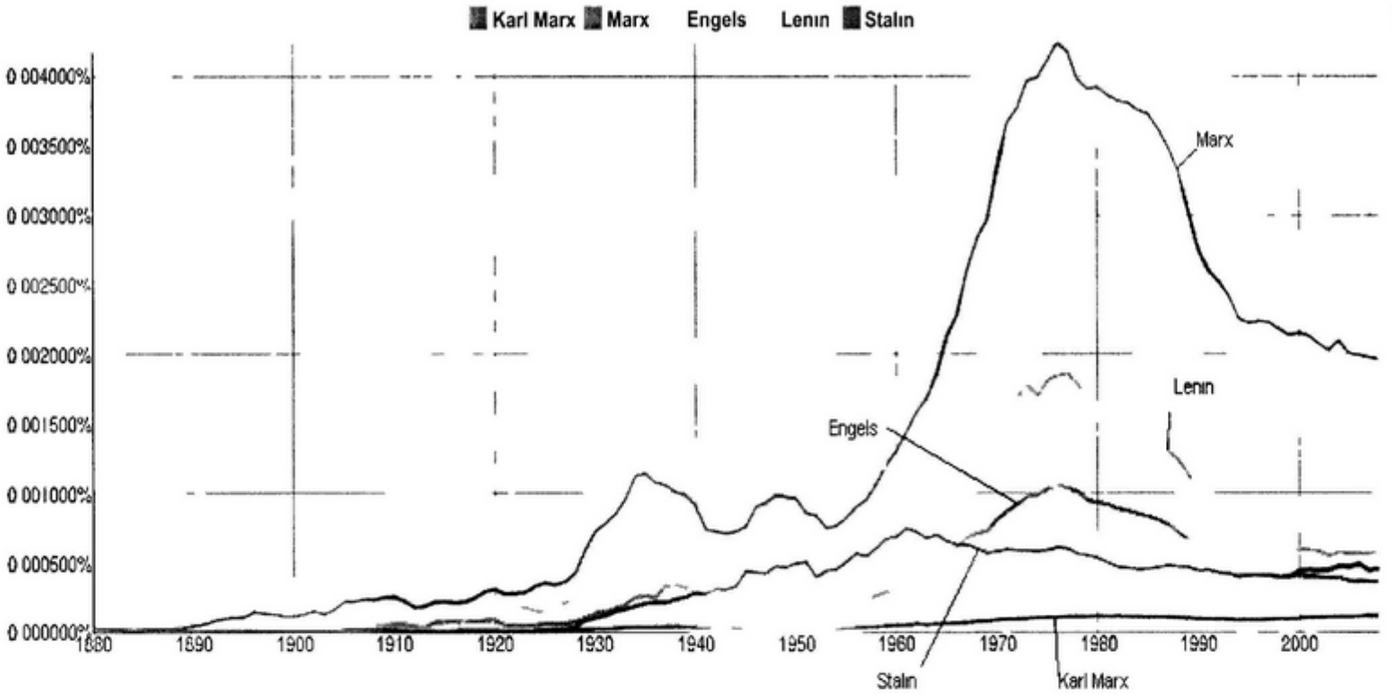

Рис 8 Нормализованные частоты встречаемости имен классиков марксизма-ленинизма в испаноязычной литературе

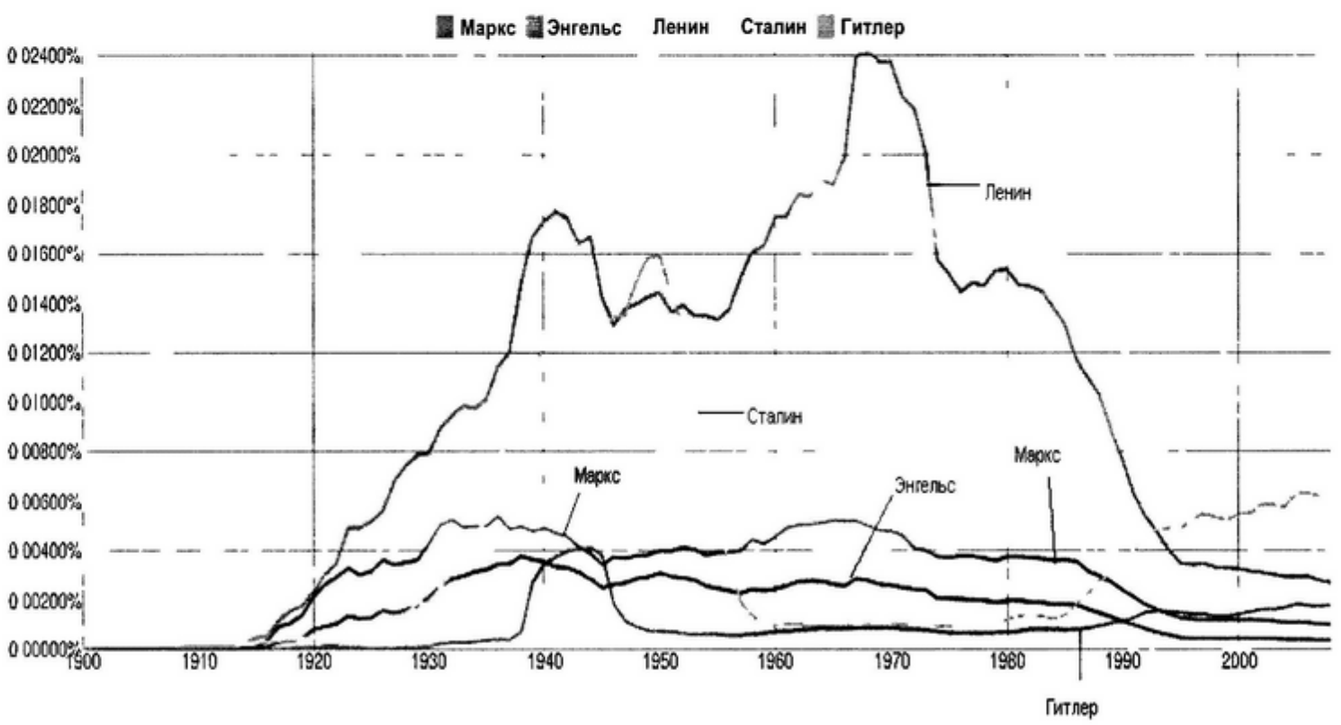

Рис 9 Нормализованные частоты встречаемости имен классиков марксизма-ленинизма в русскоязычной литературе

Таким образом, мы видим, какие уникальные возможности открываются перед гуманитариями в исследовании различных культурологических трендов на достаточно длинных промежутках времени. Эти возможности будут только улучшаться в связи с амбициозной целью компании Google – оцифровать все значимое в мире книжное наследие.

СПИСОК ЛИТЕРАТУРЫ

1. Hayes B. Bit Lit // American Scientist. – 2011. – Vol. 99, №3. – P. 190 – 194.
2. Jones E. Google Books as a General Research Collection // Library Resources & Technical Services. – 2010. – Vol. 54, №2. – P. 77 – 89.
3. Michel J.-B. et al. Quantitative analysis of culture using millions of digitized books // Science. – 2010. – Vol. 331, №1. – P. 176 – 182.

Материал поступил в редакцию 28 02 12.

Сведения об авторе

МОСКОВКИН Владимир Михайлович – доктор географических наук, профессор кафедры мировой экономики Белгородского государственного университета, профессор кафедры экологии и неоэкологии Харьковского национального университета имени В. Н. Каразина

E-mail: Moskovkin@bsu.edu.ru